UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MARCO ANTONIO INSAURRIAGA GONZALEZ

Termos e Relacionamentos em Evidência na Recuperação de Informação

Tese apresentada como requisito parcial para obtenção do grau de Doutor em Ciência da Computação

Prof. Dr. José Valdeni de Lima
Orientador

Prof[a]. Dr[a]. Vera Lúcia Strube de Lima
Co-orientadora

Porto Alegre, julho de 2005

CIP – CATALOGAÇÃO DA PUBLICAÇÃO

Gonzalez, Marco Antonio Insaurriaga

Termos e Relacionamentos em Evidência na Recuperação de Informação / Marco Antonio Insaurriaga Gonzalez — Porto Alegre: PPGC da UFRGS, 2005.

182 f.: il.

Tese (doutorado) — Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2005. Orientador: José Valdeni de Lima. Co-Orientadora: Vera Lúcia Strube de Lima.

1. Evidência. 2. Dependência de termos. 3. Recuperação de Informação. I. Lima, José Valdeni de. II. Lima, Vera Lúcia Strube de. III. Título.

AGRADECIMENTOS

Agradeço a esta experiência de doutoramento por me permitir comprovar as seguintes hipóteses:

1. Toda tese pode ser melhorada, se há orientação como a do Prof. José Valdeni de Lima e da Prof[a]. Vera Lúcia Strube de Lima.
2. Toda idéia pode ser trocada, se há colegas como Adriana Kampff, Alexandre Agustini, Ana Paula Blois, André Rodrigues, Carlos Morais, Carlos Prolo, Caroline Veraschin, Cezar Marcon, Diego de Foronda, Edicársia Pillon, Egídio Terra, Igor Steinmacher, Leonardo Langie, Luis Fernando Garcia, Luiz Pizzato, Marcírio Chaves, Maximira André, Niccholas Vidal, Pablo Gamallo, Sandra Miorelli, Tiago Telecken, Túlio Baségio e Victor Sant'Anna.
3. Toda idéia pode ser propagada e aperfeiçoada, se há alunos como Ana Martins, Daniel Müller, Daniela Toscani, Edemar Bürger, Fernanda Barão, Fernando Razera, Giovanni da Silva, Jordani Bolzoni, Letícia da Rosa, Marcéu Leite, Márcio Casado, Marcus Araújo, Maurício Oliveira, Rita Dorneles, Rodrigo Picada, Ruri Silveira e Vladimir Fraga.
4. Toda produção científica se viabiliza, se há Instituições como a Universidade Federal do Rio Grande do Sul e a Pontifícia Universidade Católica do Rio Grande do Sul.
5. Há coisas quase tão bonitas quanto o pôr-do-sol do Guaíba, como o mar do Caribe, as cordilheiras dos Andes, os glaciais da Patagônia e os castelos de Portugal.
6. Termos e relacionamentos familiares continuam em evidência, mesmo após experiência como esta, se há confiança e amor como os que tenho de minha esposa Aira e de minha filha Tatiana.

SUMÁRIO

LISTA DE ABREVIATURAS

AAj	autômato finito para nominalização de adjetivos
AEx	autômato finito para nominalização de exceções
AVb	autômato finito para nominalização de verbos
BIGR	estratégia de indexação com bigramas
CV	conjunto verbal
HMM	Hidden Markov Model
IDF	inverse document frequency
LD	lado direito
LE	lado esquerdo
LM	estratégia de indexação com lemas
NG	modelo n-grama
NILC	Núcleo Interinstitucional de Lingüística Computacional
NM	estratégia de indexação com termos nominalizados
NMRL	estratégia de indexação com termos nominalizados e RLBs
VR	estratégia de indexação com palavras originais
PLN	processamento da linguagem natural
REXTOR	Relations EXtracTOR
RLB	relação lexical binária
RI	recuperação de informação
RT	modelo relacionamento-termo
SINN	estratégia de indexação com sintagmas nominais
SN	sintagma nominal
ST	estratégia de indexação com *stems*
TCAB	trabalho correlato com árvore binária
TCBT	trabalho correlato com bitermos
TCMM	trabalho correlato com pares modificado-modificador
TCEI	trabalho correlato sobre expressões de índice
TCET	trabalho correlato com expressões ternárias
TCFM	trabalho correlato com famílias morfológicas
TCIM	trabalho correlato com índices múltiplos
TCNT	trabalho correlato com nodos temáticos
TCPL	trabalho correlato com pares lematizados
TCTR	trabalho correlato com triplas
TF.IDF	term frequency – inverse document frequency
TR	modelo termo-relacionamento
TREC	Text Retrieval Conference
TT	modelo termo-termo
UG	modelo unigrama

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Muitas abordagens para recuperação de informação (RI) assumem duas hipóteses: (i) cada termo de um documento é estatisticamente independente de todos os outros termos no texto, e (ii) métodos lingüísticos são de difícil aplicação nesta área. Contudo, há regularidades lingüísticas, produzidas pelas dependências entre termos, que precisam ser consideradas quando um texto é representado, e a representação de textos é crucial para aplicações que utilizam processamento da linguagem natural, como a RI.

Um texto é mais do que uma simples seqüência de caracteres ou palavras. As palavras apresentam características morfológicas e relações de coesão que não podem ser esquecidas na descrição dos conceitos presentes no texto. Nesse sentido, um novo modelo com dependência de termos para a RI, denominado TR+, é proposto. Ele inclui: (i) nominalização, como processo de normalização lexical, e identificação de relações lexicais binárias (RLBs) e (ii) novas fórmulas para cálculo do peso das unidades de indexação (descritores). Essas fórmulas se baseiam no conceito de evidência, que leva em conta, além da freqüência de ocorrência, os mecanismos de coesão do texto. O modelo também inclui operadores Booleanos na consulta, para complementar a especificação da dependência de termos.

Avaliações experimentais foram realizadas para demonstrar que (i) a nominalização apresenta melhores resultados em relação aos processos de normalização lexical usuais, (ii) a aquisição de informação lingüística, através de RLBs, e o uso de consultas Booleanas contribuem para a especificação de dependência de termos, e (iii) o cálculo da representatividade dos descritores baseado em evidência apresenta vantagens em relação ao cálculo baseado em freqüência de ocorrência. Os experimentos relatados indicam que esses recursos melhoram os resultados de sistemas de RI.

Palavras-chave: Recuperação de Informação, Dependência de termos, Relações Lexicais Binárias, Nominalização, Evidência.

Terms and Relationships in Evidence in Information Retrieval

ABSTRACT

Most information retrieval (IR) approaches make two assumptions: (i) each document term is statistically independent of all other terms in the text, and (ii) linguistic methods are difficult to apply in this area. However, there are linguistic regularities provided by term dependences that need to be taken into account when a text is represented, and text representation is crucial for systems that use natural language processing, such as IR.

A text is more than a simple character or word sequence. Words present morphological features and cohesion relationships, which must be considered in text concept description. In this way, a new term dependence model for IR, named TR+, is proposed. This model includes: (i) nominalization process, for lexical normalization, and binary lexical relations (BLRs) identification, and (ii) new formulas for descriptor (indexing unit) weighting based on the evidence concept, which utilizes, beside the occurrence frequency, text cohesion mechanisms. The model also includes Boolean operators in the query as a way to complement the term dependence specification.

Experimental evaluations were performed to demonstrate that (i) nominalization performs better than usual lexical normalization processes, (ii) acquisition of linguistic information, through BLRs, and the use of Boolean queries contribute to term dependence specification, and (iii) using evidence-based formula to compute descriptor representativeness presents advantages over the frequency-based one. The experiments reported in this work indicate that these resources improve the results of IR systems.

Keywords: Information Retrieval, Term Dependence, Binary Lexical Relations, Nominalization, Evidence.

1 INTRODUÇÃO

1.1 Contexto

Sistemas de recuperação de informação (RI) de documentos textuais tratam essencialmente de indexação, busca e classificação desses documentos para atender a consultas expressas por seus usuários [SAL 75, RIJ 79, SAL 83, KOW 97, SPA 97, BAE 99, MEA 2000, MOE 2000]. Um sistema de RI, em fase de indexação, constrói um índice que representa a coleção de documentos. Tal índice é constituído por unidades de indexação (chamadas descritores, neste trabalho) e seus respectivos pesos[1], formando um espaço de descritores. Em fase de busca, após ser formulada a consulta, os descritores são pesquisados. A cada documento é atribuído um valor de relevância com relação à consulta e, finalmente, os documentos são classificados por ordem decrescente desses valores.

Neste sentido, algumas questões têm preocupado os pesquisadores, das quais são destacadas nesta tese as seguintes: a seleção, a normalização e o cálculo da representatividade dos termos e os relacionamentos entre eles. Por exemplo, em relação ao texto "algumas questões têm preocupado os pesquisadores", pode ser questionado o seguinte quanto à representação: Deve-se selecionar o termo "questões", na forma original, ou gerar o termo "questão", em forma normalizada, ou o termo "quest", em forma de radical? O termo "preocupação" é representativo de algum conceito presente no texto e essa representatividade é maior, por exemplo, que a do termo "algum" (ou "algumas")? É válido o relacionamento[2] "preocupado–pesquisador"?

Desconsiderar a variação lingüística das palavras, quando da seleção dos termos, pode resultar em prejuízo à identificação de ocorrências conceitualmente próximas mas lingüisticamente diferentes [SAV 2003]. Tal identificação pode ser feita através de processos de normalização em níveis sintático, léxico-semântico ou morfológico [ARA 2000, SAV 2003]. Mais utilizada na RI, a normalização morfológica é, geralmente, realizada através de processos de conflação[3] [FRA 92, SPA 97, KOR 2004].

Quanto ao cálculo da representatividade, a indexação automática tipicamente está fundamentada na freqüência de ocorrência das palavras nos documentos [SAL 88, HIE 2000, SPA 2000], entretanto, alguns trabalhos incorporam, também, informação morfológica (e.g., [VIL 2002]) ou sintática (e.g., [LEE 2005]) para compor o espaço de descritores.

Essas questões têm sido tratadas através de três abordagens clássicas: a Booleana, já criticada há algum tempo por suas limitações inerentes (e.g., [COO 88]), como o critério binário de relevância; a vetorial, com fundamentação geométrica,

---

[1] O peso de um descritor é calculado, em relação ao documento indexado, para estabelecer a representatividade do descritor. O peso determina o quanto o descritor é representativo do texto do documento, ou seja, qual a relevância do conceito descrito por esse descritor.

[2] A expressão "relacionamento" é usada, neste trabalho, para identificar um descritor complexo, que descreve dois ou mais conceitos distintos que se relacionam de algum modo.

[3] Algoritmos de conflação (*conflation*) são aqueles que combinam a representação de duas ou mais palavras em um único termo, ou seja, reduzem variantes de uma palavra a uma forma única.

introduzida na RI por Salton e co-autores [SAL 68, SAL 75]; e a probabilística, com aplicação da teoria da probabilidade na RI demonstrada por Robertson e Sparck-Jones [ROB 76]. A abordagem vetorial tem sido hospedeira de outras abordagens para RI [SPA 2000], principalmente quanto às estratégias de indexação e, até pouco tempo, era a mais pesquisada [BAE 99]. Atualmente, em razão da forte base teórica e dos resultados apresentados [SAV 96], muitos trabalhos têm escolhido a abordagem probabilística (e.g., [LEE 2005]) e suas alternativas, como a de modelagem de linguagem (e.g., [GAO 2004]), introduzida na RI por Ponte e Croft [PON 98].

As estratégias que seguem essas abordagens podem ser divididas em dois grupos, quanto ao modelo adotado para representação de consulta e documentos: (i) as que usam modelos com unigramas e (ii) as que usam modelos com dependência de termos.

A representação com unigramas se encaixa perfeitamente nas abordagens vetorial e probabilística em virtude da assumida independência dos termos. Muitos sistemas de RI, tanto os vetoriais quanto os probabilísticos, tomam como verdadeiro que cada palavra encontrada no texto dos documentos é estatisticamente independente de todas as outras palavras. Na abordagem vetorial, esta é uma característica intrínseca ao espaço vetorial [SAL 88]. Na probabilística, a suposição de independência (*independence assumption* [ROB 94]) facilita a formalização do modelo e faz com que sua implementação se viabilize [SPA 2000].

Por outro lado, os modelos com dependências de termos se baseiam no relaxamento ou na inexistência daquela suposição. A preocupação de idealizar modelos com dependência de termos é antiga [RIJ 79, SAL 82] e persiste por duas razões: (i) é possível encontrar inconsistências na suposição de independência [COO 94], e (ii) há regularidades [LOO 2001] devidas à dependência de termos que devem ser consideradas.

Trabalhos que adotam modelos com dependência de termos fundamentam-se em dois conjuntos distintos de conhecimentos: estatísticos e lingüísticos. No primeiro caso, as dependências são geralmente identificadas através de co-ocorrência (i) por meio de n-gramas (e.g., [MIL 99]), onde um novo termo depende dos n-1 termos anteriores no texto, ou (ii) por meio de medidas estatísticas de associação de termos, como a informação mútua esperada (e.g., [LOO 2001]). Com a utilização de conhecimento lingüístico, principalmente em níveis léxico-morfológico e sintático, sintagmas nominais[4] (e.g., [LIU 2004]), pares modificado-modificador (e.g., [MAT 2000]) ou outros relacionamentos podem ser identificados apenas para indicar dependências de termos (e.g., [LEE 2005]) ou ser realmente considerados como descritores de conceitos (e.g., [VIL 2002]).

A dependência de termos tem sido especificada, então, através de diversas técnicas: cálculo de peso dos termos (e.g., [SRI 2002]), consultas com operadores Booleanos (e.g., [LOO 97]), métodos lingüísticos[5] apoiados por métodos estatísticos (e.g., [LIU 2004], ou métodos estatísticos apoiados por métodos lingüísticos (e.g., [GAO 2004]). A combinação de abordagens estatísticas, lingüísticas e Booleanas parece indicar o caminho mais promissor. Métodos lingüísticos podem ter difícil aplicação

[4] Os sintagmas nominais constituem uma classe de forma gramatical [PER 2000]. Podem apresentar comportamento sintático de sujeito, de objeto direto e, se precedidos de preposição, de adjunto adnominal e de objeto indireto.

[5] A expressão "métodos lingüísticos" é usada, neste trabalho, para identificar métodos que usam conhecimento lingüístico.

nesta área, mas recursos da área de representação do conhecimento (sub-área da inteligência artificial) são nitidamente aplicáveis à RI se a representação dos textos for enriquecida com descrições não limitadas a conjunto de termos independentes sem estrutura e cálculo de pesos formulados através de esquemas básicos [SPA 99].

1.2 Hipóteses e objetivos

No contexto descrito, o problema da RI, assumindo-se a dependência de termos, consiste em identificar os termos que representam os documentos e a consulta, normalizá-los, determinar as dependências entre esses termos, calcular a representatividade dos termos e dos relacionamentos que as dependências indicam e, finalmente, estabelecer o valor de relevância de cada documento para a consulta.

Para propor um modelo com esta orientação, são assumidas as seguintes hipóteses:

Hipótese 1: As palavras de um texto possuem determinadas características morfológicas e relações de coesão que constituem informação não negligenciável no cálculo da representatividade dos descritores delas decorrentes.

Hipótese 2: A combinação de abordagens estatísticas, lingüísticas e Booleanas produz melhores resultados do que os produzidos quando essas abordagens são adotadas isoladamente.

Hipótese 3: A aplicação de métodos que utilizam conhecimento lingüístico é viável em procedimentos automáticos de indexação, busca e classificação de documentos.

Considerando tais hipóteses, foram desenvolvidos alguns conceitos que fazem parte da presente tese: (i) nominalização, como processo de normalização lexical; (ii) relação lexical binária, como forma de identificar mecanismos de coesão do texto; e (iii) evidência, que considera tais mecanismos, além da freqüência de ocorrência de termos e relacionamentos, e contribui para o cálculo da representatividade desses descritores. Considerando tais conceitos, hipóteses mais específicas também são assumidas:

Hipótese 4: A nominalização produz resultados melhores que os obtidos com processos tradicionais de conflação.

Hipótese 5: A aquisição de informação lingüística através da identificação de relações lexicais binárias contribui para a descrição de dependências de termos.

Hipótese 6: O cálculo da representatividade dos descritores baseado em evidência melhora a classificação de relevância dos documentos, com vantagem sobre o cálculo baseado em freqüência de ocorrência.

Hipótese 7: Consultas com operadores Booleanos são uma forma eficaz de complementar a identificação de dependências de termos.

Levando em conta estas hipóteses, os objetivos deste trabalho são os seguintes:

(i) propor um novo modelo com dependência de termos para RI, que inclui:

- procedimentos automatizados para geração de termos nominalizados e identificação de relações lexicais binárias para descrever os conceitos presentes no texto dos documentos;
- cálculo de peso dos descritores baseado em evidência, para estabelecer o grau de representatividade dos mesmos quanto ao conteúdo dos documentos; e
- consulta Booleana, com formulação pré-definida, para permitir novos

critérios de classificação dos documentos por relevância.

(ii) demonstrar com resultados experimentais, através do modelo proposto e para a língua portuguesa, que as hipóteses assumidas são verdadeiras.

1.3 Organização do Texto

Após esta introdução, o texto é organizado através dos seguintes capítulos.

O capítulo 2 apresenta uma visão geral do problema da construção de espaços de descritores com termos, relacionamentos e suas representatividades. São discutidos aspectos relacionados à seleção de descritores, à normalização lingüística, ao cálculo da representatividade e à dependência de termos. Trabalhos correlatos são apresentados e analisados comparativamente.

O capítulo 3 descreve o modelo proposto. São apresentados seus componentes essenciais: a nominalização, como normalização lexical; as relações lexicais binárias, como forma de especificar dependência de termos; o conceito de evidência, adotado no cálculo da representatividade dos descritores; e a complementação do modelo através da inclusão de operadores Booleanos na consulta.

O capítulo 4 apresenta os dados disponíveis e gerados para avaliar experimentalmente o modelo proposto, assim como as ferramentas desenvolvidas e a metodologia adotada para a avaliação. A coleção de referência, utilizada neste trabalho, e as estratégias de indexação destinadas à avaliação são descritas.

Os capítulos 5 e 6 apresentam os resultados das avaliações realizadas. O capítulo 5 diz respeito aos resultados obtidos pelos procedimentos executados para processamento da linguagem natural. O capítulo 6 trata dos resultados de recuperação obtidos pela implementação do modelo proposto, analisados comparativamente com os resultados de outras estratégias.

O capítulo 7 analisa a relação custo/benefício da proposta e apresenta a análise de complexidade dos algoritmos utilizados.

No Capítulo 8 são discutidos aspectos importantes relacionados à pesquisa desenvolvida e à experiência realizada para a avaliação do modelo. São destacadas as principais diferenças entre a proposta e outros modelos. Causas de falhas na recuperação são levantadas, propriedades dos espaços de descritores são discutidas e características que beneficiam uma estratégia de RI são sugeridas.

O capítulo 9 resume as contribuições da tese, indica trabalhos futuros e informa sobre as publicações decorrentes deste trabalho.

2 TERMOS, RELACIONAMENTOS E REPRESENTATIVIDADE

2.1 Introdução

Indexação de textos é o processo que estabelece descritores dos conteúdos dos textos de uma coleção de documentos, com propósito de busca e classificação dos mesmos para atender consultas em sistemas de RI. Descritores[6] podem descrever conceitos atômicos, sendo termos, ou conceitos complexos, sendo relacionamentos. Cada descritor pode ser mais ou menos representativo do texto de um documento e essa variação deve ser conhecida.

O conjunto de descritores gerado na indexação provê uma visão lógica dos documentos [BAE 99], com o objetivo de associar esses descritores a conceitos presentes nos textos dos documentos. Os descritores e seus respectivos graus de representatividade formam um espaço de descritores, definido a seguir.

Definição: Espaço de descritores

Dados um conjunto de conceitos $C = \{c\}$, uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I é definido como:

$$I = T \cup R$$

onde:

$T = \{t, t$ é um termo que descreve $c \in C$ e tem peso $W_{t,d}$ em $d \in D\}$ e

$R = \{r, r$ é um relacionamento que representa uma relação de dependência entre $t_1 \in T$ e $t_2 \in T$, para descrever conceito formado pela combinação de $c_1 \in C$ e $c_2 \in C$, e tem peso $W_{r,d}$ em $d \in D\}$.

Para que um espaço de descritores seja construído, o texto dos documentos deve ser pré-processado de acordo com a estratégia adotada. São usuais os procedimentos de *toquenização*[7], seleção de descritores e conflação [BAE 99]. Se a estratégia adotada necessita de informação lingüística, outros métodos devem ser considerados. A análise morfo-sintática do texto pode ser obtida através de um *parser* [MAR 2003]. Podem ser incluídas, também, anotações para resolução de co-referências [VIE 2002]. Essas informações são, geralmente, inseridas através da etiquetagem[8] de texto após a *toquenização*.

[6] Será preferida, neste trabalho, a expressão "descritor" à expressão "termo de índice" por ser mais adequada à identificação das unidades de indexação, pois estas podem ser constituídas por termos ou por relacionamentos. Assume-se que todo descritor possui pesos (binários ou não), e cada peso determina a representatividade do descritor em relação a um texto.

[7] Identificação de cada item lexical do texto, tal como palavra e pontuação.

[8] Um etiquetador gramatical (*part-of-speech tagger*) identifica, com a colocação de uma etiqueta (*tag*), a categoria gramatical de cada palavra do texto. Geralmente é morfológico (identifica somente a categoria morfológica) ou morfo-sintático (identifica também as funções sintáticas).

Após o pré-processamento do texto dos documentos, a geração do espaço de descritores, propriamente dita, acontece. Nesse processo se dá a seleção e a normalização (se houver) dos descritores e, ainda, a obtenção das informações necessárias (geralmente a freqüência de ocorrência) para o cálculo de seus pesos. Tal cálculo avalia a representatividade dos descritores em relação ao conteúdo do texto e à relevância dos conceitos descritos em cada documento. O peso de um descritor é, entretanto, afetado, indiretamente, pelas estratégias adotadas para sua seleção e normalização, conforme é discutido a seguir.

2.2 Termos

A expressão "termo" é entendida aqui como uma unidade lexical formada por uma única palavra ou por mais de uma. Neste sentido, a expressão "termo composto" tem sido usada para indicar uma unidade lexical complexa com características de uma única unidade sintática e semântica [DIA 2000], descrevendo um único conceito [STR 96]. Por exemplo, "boca da noite" será um termo, se for identificado como uma única unidade lexical, ou, ao contrário, constituirá mais de um termo. Neste último caso, conforme a abordagem, poderá ser reconhecida ou não uma dependência entre os termos "boca" e "noite". Termos (compostos ou não) são, de qualquer forma, descritores.

Salton e MacGill utilizam o gráfico de Luhn [SAL 83] (Figura 2.1) para definir os bons descritores e discutir como obtê-los. Segundo este gráfico, a expressividade dos termos com freqüência de ocorrência intermediária é maior que a dos termos muito ou pouco freqüentes.

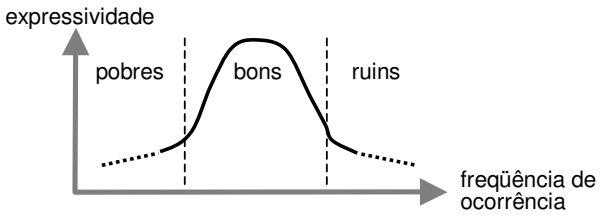

Figura 2.1: O gráfico de Luhn e a classificação dos descritores

Bons descritores deveriam ser, então, derivados por transformação dos termos pobres em termos bons através de relações encontradas em um thesaurus[9], ou por transformação dos termos ruins através da construção de frases.

A usual eliminação de s*topwords* faz parte do processo de seleção de descritores. *Stopwords* são palavras como preposições, artigos e conjunções, que podem ser descartadas em fase de indexação e na consulta. Essa eliminação, com conseqüente redução do espaço de descritores, justifica-se porque as *stopwords* são consideradas desnecessárias como descritores, por terem pouca representatividade. São tidas como descritores ruins, ou seja, com pouca expressividade e alta freqüência de ocorrência.

Ao serem descartadas as preposições, entretanto, perde-se algum significado.

[9] Thesauri são dicionários que não definem as palavras, como tradicionalmente é feito, mas as organizam como conceitos relacionados semanticamente entre si.

Por exemplo: "caixa de vidro" e "caixa para vidro" têm, evidentemente, significados diferentes e esta diferença não pode ser representada, neste caso, sem as preposições. Desta forma, ganha-se em economia mas perde-se em representatividade.

Não só a normalização léxico-semântica (com thesaurus), através de sinonímia, por exemplo, possibilita a transformação de descritores pobres em descritores bons. Processos para normalização lexical (conforme é discutido na próxima Seção) também geram descritores com maior expressividade, pois as freqüências das formas normalizadas serão maiores que as dos termos originais, mais raros por suas variações morfológicas.

Por outro lado, há descritores originalmente pobres que não conseguem aumentar sua expressividade, mesmo através da normalização, levando em conta apenas a freqüência de ocorrência. Entretanto, são bons representantes de conceitos relevantes presentes no texto. Neles o autor pode estar centralizando uma idéia importante que pretende comunicar, ainda que não concretize isso, essencialmente, através da freqüência com que os usa. Por exemplo, um substantivo com a função de núcleo do sujeito de uma oração terá maior evidência que se utilizado como adjunto adnominal. Tal evidência, portanto, não depende apenas da freqüência de ocorrência.

2.2.1 Normalização lingüística

A seleção dos descritores, a quantidade dos mesmos e o peso de cada um podem ser afetados pela estratégia adotada para normalização lingüística. O reconhecimento de variações lingüísticas encontradas em um texto permite o controle de vocabulário [JAC 97]. Tal controle determina termos preferenciais a serem usados como descritores, ou seja, os seleciona, os restringe em número e, em conseqüência, influencia o cálculo da representatividade dos mesmos.

Há três tipos de normalização lingüística [ARA 2000, SAV 2003]: sintática, léxico-semântica e morfológica (ver Tabela 2.1).

A normalização sintática ocorre quando há a transformação de frases semanticamente equivalentes mas sintaticamente diferentes, em uma forma única e representativa das mesmas, como "eficiente processo rápido" e "processo rápido e eficiente", que poderiam ter uma representação comum.

A normalização léxico-semântica ocorre quando são utilizados relacionamentos semânticos (como a sinonímia) para substituir palavras morfologicamente distintas por uma única forma que representa o conceito referenciado.

Tabela 2.1: Normalização lingüística

normalização		método usual
lexical	morfológica	conflação
	léxico-semântica	busca de relações semânticas em thesaurus
sintática		aplicação de regras gramaticais

Em nível lexical há dois extremos de normalização. De um lado há a normalização léxico-semântica, através de busca de sinônimos, por exemplo, em thesaurus, considerando informações terminológicas [JAC 99]. Em outro extremo está a normalização morfológica, através da conflação, que explora similaridades morfológicas inferindo proximidades conceituais.

A normalização morfológica ocorre quando há redução das formas flexionais de

uma palavra, através de conflação, a uma forma única que procura representar um conceito ou uma classe de conceitos. Os processos mais comuns de conflação são o *stemming* [FRA 92, KRO 93, ALL 2003] e a lematização [ARA 2000, KOR 2004]. S*temming* é um processo que reduz ao mesmo *stem* (parte fundamental semelhante ao radical) palavras que se diferenciam basicamente pela flexão, como:

stemming(livro) = *stemming*(livros) = livr ou

stemming(caminhada) = *stemming*(caminhei) = caminh.

A lematização reduz as palavras variáveis à correspondente forma canônica: verbos no infinitivo e palavras, como substantivos e adjetivos, no singular e, quando existir, masculino. São exemplos:

lematização(livro) = lematização(livrinho) = livro,

lematização(livre) = lematização(livres) = livre ou

lematização(caminhar) = lematização(caminhei) = caminhar.

A principal diferença entre os resultados de *stemming* e de lematização é que, no primeiro caso, palavras de diferentes categorias morfológicas podem ter o mesmo *stem*, como:

stemming(construiu) = *stemming*(construções) = constru,

enquanto que, na lematização, a categoria morfológica é mantida:

lematização(construiu) = construir ≠ lematização(construções) = construção.

Conforme Braschleer e Ripplinger [BRA 99], os benefícios do *stemming* são reconhecidos na RI. Este processo (i) reduz o número de descritores e o tamanho do arquivo de índice, e (ii) torna a recuperação independente da forma com que o termo ocorre na consulta. O mesmo pode ser dito sobre a lematização. Isso confirma que os processos de conflação têm um segundo efeito, além da normalização lexical em si: a economia na quantidade de descritores e no espaço de memória necessário para armazená-los.

Apesar dos benefícios desses métodos de conflação, alguns problemas ainda estão por ser resolvidos. O *stemming*, por exemplo, não tem sucesso com termos onde a flexão é raramente usada ou inexiste (por exemplo, nomes próprios) [BRA 99]. A análise flexional e morfológica de termos compostos também é problemática, mesmo para a língua Inglesa [SAV 2003]. Uma solução, nesses casos, é a decomposição do termo e a aplicação da normalização, separadamente, a cada componente, especialmente na lematização [KOR 2004].

Alguns significados podem ser perdidos no *stemming* e palavras de famílias de significados diferentes podem ser agrupadas. Esses erros podem ser difíceis de detectar e corrigir em sistemas automáticos, requerendo esforço adicional para tratamento de exceções. Seriam os casos de:

stemming(livro) = *stemming*(livre) = livr e

stemming(caminhada) = *stemming*(caminhão) = caminh.

Enquanto um algoritmo simples de *stemming* é suficiente para Inglês, estratégias mais sofisticadas são necessárias para idiomas com morfologia flexional complexa, como Alemão, Espanhol, Finlandês, Francês e Português. *Stemmers* para tais idiomas apresentam elevado custo computacional [VIL 2002].

Experimentos com Espanhol [VIL 2002] e Finlandês [KOR 2004] concluem que a lematização produz, na RI, melhores resultados que o *stemming*. Entretanto, a lematização também apresenta problemas. Algumas palavras pertencentes à mesma

família de significados podem não ser normalizadas, como:

lematização(livre) = livre ≠ lematização(liberdade) = liberdade e

lematização(caminhei) = caminhar ≠ lematização(caminhada) = caminhada.

2.3 Cálculo da representatividade

Após a seleção e a normalização dos descritores, é necessário calcular a representatividade, que é uma das propriedades básicas de um descritor. O grau de representatividade de um descritor *i* em um documento *d* é dado pelo peso $W_{i,d}$. Na abordagem vetorial, um dos modos usuais de calcular esse peso é o seguinte [SAL 88]:

$$W_{i,d} = \frac{w_{i,d} IDF_i}{\sqrt{\sum_j \left(w_{j,d} IDF_j\right)^2}} \quad (1)$$

onde:

j representa cada um dos descritores do espaço de descritores;

$w_{j,d} = f_{j,d}$ ou $w_{j,d} = \frac{f_{j,d}}{\max f_d}$ (o mesmo vale para $w_{i,d}$);

$f_{j,d}$ = freqüência de ocorrência de *j* em *d*;

max f_d = máxima freqüência de ocorrência dos descritores de *d*;

$IDFj = \log \frac{N}{df_j}$ (o mesmo vale para IDF_i);

N = número de documentos na coleção; e

df_j = número de documentos onde *j* ocorre.

No modelo probabilístico, um dos esquemas usuais de cálculo do peso de um descritor *i* em um documento *d* corresponde à fórmula Okapi BM25 [ROB 94, SPA 2000]:

$$W_{i,d} = \frac{w_{i,d}(k_1 + 1)}{k_1((1-b) + b\frac{DL_d}{AVDL}) + w_{i,d}} IDF_i \quad (2)$$

onde:

$w_{i,d} = f_{i,d}$ é a freqüência de ocorrência de *i* em *d*;

k_1 e *b* são parâmetros (discutidos adiante);

DL_d é o comprimento (quantidade de palavras) do documento *d*;

AVDL é o comprimento médio dos documentos da coleção; e

IDF_i é o mesmo fator utilizado na Equação 1.

O fator *IDF* (*inverse document frequency*) é utilizado para penalizar descritores que ocorrem com muita freqüência na coleção [HIE 2000]. O uso da freqüência de ocorrência e do fator *IDF* caracteriza esquemas conhecidos como TF.IDF [SAL 88, HIE 2000].

O parâmetro k_1 (da Equação 2) é utilizado para correção da freqüência. Se $k_1 = 0$, o peso será binário; se k_1 assumir valores elevados, a relação entre o peso e $f_{i,d}$ será aproximadamente linear; e se k_1 assumir valores levemente superiores a 1, esta relação será altamente não linear. Com valores de k_1 entre 1,2 e 2 (valores usuais – especialmente 1,2), ocorrências adicionais do descritor, acima da terceira ou quarta

ocorrência em um documento, têm impacto mínimo no cálculo [SPA 2000].

A utilização do parâmetro *b* (da Equação 2) e do comprimento (quantidade de palavras) do documento está relacionada às hipóteses do escopo e da verbosidade [ROB 94]. Na hipótese do escopo, documentos mais longos têm mais informação que documentos menos longos. Na hipótese da verbosidade, documentos mais longos possuem escopo similar ao de um documento menos longo, simplesmente usam mais palavras.

Os efeitos destas hipóteses sobre o cálculo da representatividade dos descritores são antagônicos. Com a hipótese do escopo fica justificado o cálculo da importância dos descritores independente do comprimento dos documentos. Com a hipótese da verbosidade, ao contrário, documentos mais longos devem ser penalizados. Na prática, documentos de coleções reais combinam estes dois efeitos [ROB 94]. Com relação ao parâmetro *b* da Equação 2, se $b = 1$, anula-se a hipótese do escopo predominando somente a da verbosidade; se *b* assumir valores menores que 1, diminui a importância da verbosidade; e se $b = 0$, a hipótese da verbosidade se anula. Valores usuais ficam em torno de 0,75 [SPA 2000].

2.4 Relacionamentos

Até aqui, foram discutidos a seleção, a normalização e o cálculo da representatividade sempre tendo em vista descritores constituídos como termos. Entretanto, um espaço de descritores, conforme a abordagem adotada, pode incluir também os relacionamentos entre esses termos.

Alguns sistemas de RI permitem consultas através de frases, como "defesa eficiente" e "feira de domingo". Nestes casos, o usuário não estaria interessado em qualquer "defesa" ou em qualquer "feira". A coincidência exata poderia garantir, assim, o atendimento da consulta. Entretanto, nem sempre a coincidência exata é satisfatória. Para os exemplos citados, a expressão "defender eficientemente" poderia ser relevante para a consulta "defesa eficiente". Da mesma forma, "feira dominical" poderia interessar ao usuário da consulta "feira de domingo". Isto ocorre porque construções sintaticamente diferentes podem ter o mesmo significado.

Como representar esses significados através de relacionamentos? Por exemplo, o trecho "... têm preocupado os pesquisadores" pode ser representado através de um dos seguintes tipos de relacionamentos: o par modificado-modificador "pesquisador–preocupado", o bigrama "(preocupado,pesquisador)", o sintagma nominal "pesquisador preocupado", ou algum outro formato, como a expressão ternária "preocupação-de-pesquisador" e a relação binária "de(preocupação,pesquisador)". Dois, dentre esses tipos de relacionamentos, têm características especiais: os bigramas e os sintagmas nominais.

Há duas diferenças importantes entre a descrição produzida pelos bigramas e a de outros tipos de relacionamentos. A primeira diferença é a inviável descrição de conceito pretendida através de alguns bigramas, como "(ferro,sopa)", capturado de "panela de ferro com sopa", por exemplo. A segunda diferença é que, quando a descrição é viável, bigramas como "(a,x)" e "(b,x)" representam acepções diferentes do descritor "x": uma, quando antecedido por "a", outra, por "b", cada uma com sua representatividade. Por exemplo, nos bigramas "(sentar,banco)" e "(depositar,banco)", o termo "banco" produz descrições diferentes em cada um deles, sendo os termos "sentar" e "depositar" (importantes) coadjuvantes.

Sintagmas nominais constituem outro caso especial de tipo de relacionamento porque podem representar tanto conceitos complexos quanto atômicos. Conceitos

atômicos podem ser descritos por sintagmas nominais formados por um único substantivo, como "noite", ou mesmo por mais de uma palavra, como "boca da noite". O sintagma nominal "boca da noite" poderá ser um termo (composto), conforme já foi mencionado, ou um relacionamento. De qualquer forma será um descritor.

Tendo sido estabelecidos os formatos para representar dependências de termos, como identificá-las? Diversos pesquisadores têm analisado a aplicação de conhecimentos estatísticos e lingüísticos no tratamento desse problema. Conforme Salton [SAL 86], quando relacionamentos baseados em estatística são gerados, um grande número deles tem utilidade, outros tantos não. Alternativas válidas não existiam ainda segundo Salton, que justifica afirmando que os resultados de métodos baseados em sintaxe não eram na época encorajadores. Entretanto, Fagan [FAG 87] analisou a representação de relacionamentos estatísticos e sintáticos e apontou vantagens para a segunda abordagem, como a capacidade de identificar relacionamentos entre palavras não adjacentes, como entre as palavras "algoritmos" e "concorrentes" em "algoritmos seqüenciais e concorrentes", por exemplo.

Há ainda outras questões em aberto, além da identificação das dependências entre termos. Por exemplo, é preciso decidir como a representatividade dos relacionamentos deve ser calculada e como esses descritores devem ser normalizados. O encaminhamento dessas questões passa pela especificação completa do espaço de descritores.

2.5 Modelos de espaços de descritores

A partir da análise das estratégias para RI que consideram dependência de termos, encontradas na bibliografia, é possível classificar os espaços de descritores construídos por elas em quatro modelos de representação de texto alternativos[10]: n-grama (NG), termo-termo (TT), termo-relacionamento (TR) e relacionamento-termo (RT). Essa classificação, com ênfase na definição dos espaços de descritores e incluindo o modelo com unigramas (UG), é apresentada na Figura 2.2.

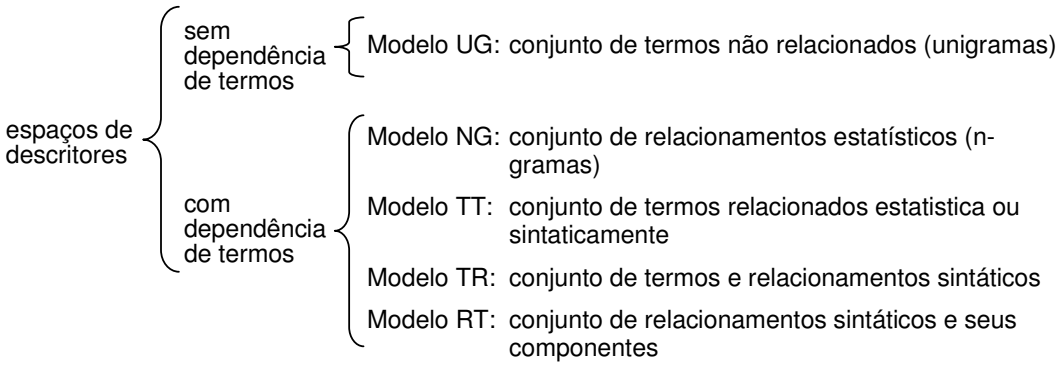

Figura 2.2: Classificação dos espaços de descritores

[10] Esta classificação se inspira em aspectos daquela apresentada por Croft, Turtle e Lewis, para modelos de dependência de termos [CRO 91]. Entre outras diferenças, esses autores sugerem um modelo onde termos e frases constituem descritores independentes. Conforme os critérios adotados aqui, tais estratégias seguiriam o modelo com unigramas, sendo as frases consideradas como termos compostos.

Conforme a Figura 2.2, no modelo UG, a representação de um texto é concebida como um conjunto de termos não relacionados entre si. O espaço de descritores neste modelo é definido a seguir.

Definição: Espaço de descritores no modelo UG

Dados uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{UG} é definido como:

$$I_{UG} = T ,$$

sendo que:

se $t \in T$ ocorre em $d \in D$, então a representatividade de t depende principalmente da sua freqüência de ocorrência em d. R é vazio.

No modelo NG, a representação de um texto é concebida como um conjunto de relacionamentos estatísticos entre termos. O uso de n-gramas (especialmente bigramas) caracteriza este modelo. O espaço de descritores neste modelo é definido a seguir.

Definição: Espaço de descritores no modelo NG

Dados uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{NG} é definido como:

$$I_{NG} = R ,$$

sendo que:

se $t_1 \in T$ está relacionado a $t_2 \in T$ através de $r \in R$ em $d \in D$, então a representatividade de r depende da freqüência de co-ocorrência de t_1 e t_2 em d.

No modelo TT, a representação de um texto é concebida como um conjunto de termos relacionados estatística ou sintaticamente. Os relacionamentos não fazem parte do espaço de descritores e podem ser estabelecidos através de medidas estatísticas de associação de termos ou por métodos lingüísticos. Os relacionamentos formam principalmente pares modificado-modificador. O espaço de descritores neste modelo é definido a seguir.

Definição: Espaço de descritores no modelo TT

Dados uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{TT} é definido como:

$$I_{TT} = T ,$$

sendo que:

se $t_1 \in T$ está relacionado a $t_2 \in T$ através de $r \in R$ em $d \in D$, então a representatividade de t_1 depende da sua própria freqüência de ocorrência, mas também é afetada por t_2, em co-ocorrência ou não com t_1.

No modelo TR, a representação de um texto é concebida como um conjunto de termos e de relacionamentos sintáticos. Pares modificado-modificador constituem os principais relacionamentos neste modelo. Métodos lingüísticos, para a identificação dos relacionamentos, podem ser apoiados por métodos estatísticos. O espaço de descritores neste modelo é definido a seguir.

Definição: Espaço de descritores no modelo TR

Dados uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{TR} é definido como:

$$I_{TR} = R$$

ou

$$I_{TR} = T \cup R ,$$

sendo que:

> se $t_1 \in T$ está relacionado a $t_2 \in T$ através de $r \in R$ em $d \in D$, então a representatividade de r depende da freqüência de ocorrência de t_1 e de t_2 em d. Se T é incluído em I_{TR}, então as representatividades de t_1 e t_2 dependem das suas próprias freqüências de ocorrência em d.

No modelo RT, a representação de um texto é concebida como um conjunto de relacionamentos sintáticos e seus componentes. Expressões ou sintagmas (especialmente nominais) são utilizados como principais descritores neste modelo. Métodos lingüísticos, para a identificação dos relacionamentos, podem ser apoiados por métodos estatísticos. O espaço de descritores neste modelo é definido a seguir.

Definição: Espaço de descritores no modelo RT

Dados uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{RT} é definido como:

$$I_{RT} = R$$

ou

$$I_{RT} = T \cup R ,$$

sendo que:

> a representatividade de $r \in R$ em $d \in D$ depende da sua própria freqüência de ocorrência em d. Se T é incluído em I_{RT} e se $t \in T$ é componente de r, então a representatividade de t depende da representatividade de r.

No próximo Capítulo é proposto o novo modelo TR+, que combina aspectos dos modelos TR e RT, quanto ao cálculo do peso dos descritores. A proposta inclui um processo alternativo de normalização lexical, tipos próprios de relacionamentos entre termos e o conceito de evidência, visando melhorar a representatividade do espaço de descritores. Antes, estratégias com dependência de termos são apresentadas a seguir, como trabalhos correlatos que exemplificam os modelos descritos nesta Seção.

2.6 Trabalhos correlatos

O modelo proposto nesta tese resulta da combinação de estratégias distintas. Por essa razão, diversos experimentos, que se encaixam no modelo com dependência de termos TR, apresentam aspectos em comum com a proposta. Além desses casos, também são de interesse experiências que podem ser comparadas com a proposta, mesmo adotando o modelo RT, ou ainda os modelos NG e TT. Dez trabalhos são detalhados nesta Seção. Eles foram selecionados de acordo com critérios que levam em conta recentidade, quantidade de informação disponível e semelhança com a presente proposta.

Diversos trabalhos foram considerados para a formulação dos modelos

apresentados na Seção anterior. Alguns deles foram examinados com maior interesse, embora não estejam detalhados nesta Seção em razão de algum dos critérios mencionados anteriormente. Tais trabalhos, entretanto, são bons exemplos para os modelos NG, TT, TR e RT:

- De acordo com o modelo NG, Song e Croft [SON 99] incorporam dependência de termos através do uso de bigramas. De acordo com o mesmo modelo, Miller, Leek e Schwartz [MIL 99] usam *Hidden Markov Model* (HMM) para implementar dependência de termos também através de bigramas.
- Segundo o modelo TT, Gao, Nie, Wu e Cao [GAO 2004] apresentam um método para identificação de dependências de termos usando uma gramática e uma estrutura em forma de grafo. Essa estrutura limita as dependências aos relacionamentos identificados pela gramática.
- Seguindo o modelo TR, Maarek e Smadja [MAA 89] identificam pares modificado-modificador (como verbo-objeto direto/indireto e substantivo-adjetivo) e, para calcular a representatividade desses relacionamentos, adotam a fórmula $(-\log_2(P_{i,d}P_{j,d}))f_{ij,d}$, onde $f_{ij,d}$ é a freqüência de ocorrência do par e $P_{i,d}$ e $P_{j,d}$ são, respectivamente, as probabilidades de ocorrência do termo modificado *i* e do termo modificador *j* em um documento *d*.
- Conforme o modelo RT, Zhai [ZHA 97] usa um analisador sintático para sintagmas nominais em fase de indexação e combina três diferentes tipos de descritores extraídos de cada sintagma nominal identificado: termos, pares modificado-modificador e o próprio sintagma nominal. Também seguindo o modelo RT, Liu e co-autores [LIU 2004] identificam sintagmas nominais na consulta e recuperam documentos tomando como base essa identificação. Essa estratégia considera que um documento *d* é relevante para uma consulta com um sintagma nominal *sn* se todos os termos de *sn* ocorrem no interior de uma janela de texto de tamanho predeterminado em *d*.

A seguir são, então, descritos os trabalhos correlatos selecionados e, ao final deste Capítulo, é apresentada uma análise comparativa dos mesmos.

2.6.1 Expressões de índice

Wondergem, Bruza e co-autores [BRU 91, WON 2000, WON 2000a] descrevem o que chamam de "expressões de índice". Tais expressões podem caracterizar o conteúdo de um texto através de construções denominadas "lithoids". As expressões de índice apresentam o seguinte padrão, descrito utilizando-se formalismo EBNF:

termo { *conector termo* }*

Essas expressões estão relacionadas a sintagmas nominais, sendo que: (i) um termo pode ser uma palavra-chave, um conceito, uma denotação ou um valor de atributo; e (ii) um conector, que representa uma relação entre termos, pode ser uma preposição, um verbo no gerúndio ou o conector nulo, representado por "**.**" (ponto). Alguns tipos de relacionamentos, representados principalmente através de preposições, são: posse ou ação-objeto (*of*), ação-agente (*by*), posição (*in*, *on*, ...), associação direta (*to*, *on*, *for* e *in*), associação (*with* e *and*) e equivalência (*as*).

A derivação de expressões de índice a partir de um texto é realizada através da construção de árvores de representação de sintagmas nominais como, por exemplo, "a poluição da água por metais" (ver Figura 2.3). Após a remoção de *stopwords* (com

exceção das preposições), os termos remanescentes são sucessivamente processados. No caso de ser encontrado um conector, ele servirá de guia para decidir se a árvore corrente será aprofundada ou alargada.

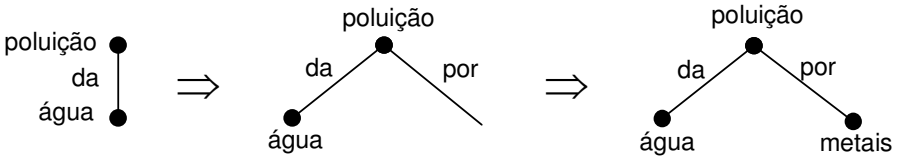

Figura 2.3: Exemplo de derivação de uma expressão de índice

Um "lithoid" (Figura 2.4 – [WON 2000a]) pode ser, então, construído a partir da árvore de representação (Figura 2.3 – adaptada de [BRU 91]). Um "lithoid", denominado assim em razão de sua estrutura semelhante à de um cristal, é um grafo cujos nodos são constituídos pela expressão de índice inicial e por todas as suas subexpressões. Os arcos conectam cada subexpressão com sua expressão correspondente. O símbolo ε representa uma expressão de índice vazia, que é subexpressão válida para qualquer expressão.

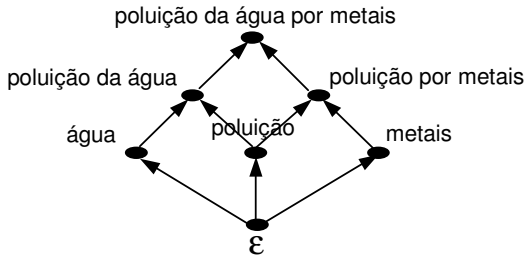

Figura 2.4: Exemplo de "lithoid"

Os nodos de um "lithoid" são vinculados ao texto onde a correspondente expressão ocorre. Assim, alguns nodos podem estar vinculados e outros não. No caso da Figura 2.4, se os nodos "poluição por metais", "poluição" e "metais" estiverem vinculados a um determinado texto, isto significaria que ele trata de temas associados a estas expressões. O mesmo texto pode não estar vinculado ao nodo "água", por exemplo.

Uma aplicação para RI foi desenvolvida, sendo usada uma estrutura chamada Arquitetura de Índice de Associação [WON 2000], construída a partir de expressões de índice. Nela, os "lithoids" são utilizados para permitir consultas por navegação.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Wondergem, Bruza e co-autores, deste trabalho correlato sobre expressões de índice (TCEI), podem ser destacados os seguintes aspectos:

- Ambas as estratégias usam termos e relacionamentos como descritores, embora, nas publicações disponíveis, não seja mencionado como TCEI realiza o cálculo de peso para os descritores.
- No modelo TR+ é usada nominalização como processo de normalização lexical. Nas publicações disponíveis sobre TCEI, os exemplos são apresentados com termos aparentemente lematizados.
- Ambas as estratégias usam preposições como identificadores de relacionamentos, embora no modelo TR+ elas não sejam rotuladas, como

ocorre em TCEI.

2.6.2 Índices múltiplos

Strzalkowski e co-autores [STR 96, STR 99] propõem um sistema onde o texto dos documentos é inicialmente processado através de um analisador sintático. Uma gramática com 400 regras é aplicada sentença a sentença. Relacionamentos são extraídos da árvore de análise sintática construída. Eles são usados como descritores juntamente com os termos. Nessa extração, é considerada a distribuição estatística dos componentes para decidir se a associação entre dois termos é sintaticamente válida e semanticamente significante.

Stopwords (preposições, conjunções, artigos, etc.) são removidas do texto. O *stemmer* morfológico tradicional é substituído por um eliminador de sufixos assistido por dicionário que (i) reduz variantes de palavras a suas respectivas raízes, conforme dicionário, e (ii) converte substantivos derivados de formas verbais (tais como os substantivos, em Inglês, "implementation" e "storage") nas raízes dos verbos correspondentes (respectivamente "implement" e "store"), também com o auxílio de dicionário.

Nomes próprios são identificados e representados como termos. Os relacionamentos extraídos da árvore de análise sintática são pares modificado-modificador. Esses pares são normalizados sintaticamente. Por exemplo (em Inglês): formas como "weapon profileration", "proliferation of weapons" e "proliferate weapons" são reduzidas à forma "weapon+proliferate". Termos compostos (como "joint venture") podem ser extraídos, no lugar de relacionamentos, de acordo com a distribuição estatística dos componentes de cada par analisado.

São gerados, ao final, quatro arquivos de índices: de termos, de termos compostos, de nomes próprios e de relacionamentos. É concebido um fluxo de busca onde descritores são comparados com a consulta em quatro etapas. Cada um dos arquivos de índices é acessado em cada etapa. Após, os resultados parciais são combinados para estabelecer a classificação dos documentos.

É utilizado o esquema TF.IDF para todos os tipos de descritores. Strzalkowski e co-autores acreditam que este esquema não é apropriado para descritores de diferentes tipos (como termos simples, nomes próprios e relacionamentos). Eles suspeitam que todos os esquemas de cálculo de pesos baseados em freqüência de ocorrência apresentem esta desvantagem. Por esta razão, são introduzidos acréscimos aos pesos dos relacionamentos, através de parâmetros que representam fatores de multiplicação.

Strzalkowski e co-autores também acreditam que abordagens com inclusão de conhecimento lingüístico apresentam vantagem sobre aquelas puramente estatísticas. Ressalvam, entretanto, que essas abordagens serão mais efetivas se atenderem aos distintos níveis de representação que os textos exigem.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Strzalkowski e co-autores, deste trabalho correlato com índices múltiplos (TCIM), podem ser destacados os seguintes aspectos:

- Ambas as estratégias usam termos e relacionamentos como descritores.
- Em TCIM é usado um processo semelhante a *stemming*, como normalização morfológica, enquanto no modelo TR+ é usada nominalização.
- Ambas as estratégias usam regras gramaticais para a identificação de relacionamentos, incluindo normalização sintática.

- Em TCIM, o cálculo do peso dos descritores é realizado através de esquema TF.IDF e, no caso dos relacionamentos, há inclusão de parâmetros para ajustes. Por outro lado, no modelo TR+ tal cálculo se baseia no conceito de evidência.
- Ambas as estratégias usam diversos arquivos de índices, ao contrário do que normalmente ocorre com outros trabalhos.

2.6.3 Nodos temáticos

Loukachevitch, Dobrov e co-autores [DOB 98, LOU 99, LOU 2000] propõem a representação temática de textos através de estrutura hierárquica de nodos temáticos (ou conceituais), para RI e para sumarização automática. A abordagem pressupõe que aquelas cadeias lexicais que caracterizam o tema principal de um texto, normalmente, possuem elementos que ocorrem juntos nas sentenças com maior freqüência que outros elementos de outras cadeias lexicais. É assumido o seguinte conjunto de premissas:

- a coesão de um texto pode ser obtida através de referências, elipses, conjunções e termos semanticamente relacionados;
- a coesão lexical é o tipo mais freqüente de coesão textual, e pode ocorrer através de repetições, de relacionamentos como sinonímia e hiponímia, ou de relações sintagmáticas;
- cadeias lexicais (seqüências de termos conectados através de coesão lexical) estão extremamente relacionadas à estrutura temática do texto, sendo, portanto, de importância crucial para o processamento e para a representação do conteúdo tratado;
- a coesão lexical não está baseada em um conjunto isolado de cadeias lexicais mas em uma rede complexa constituída por relações diversas entre os termos; e
- o tema de um texto pode usualmente ser descrito através de temas menos gerais que, por sua vez, podem ser caracterizados por temas ainda mais específicos, e assim por diante, ou seja, a representação temática de um texto é uma estrutura hierárquica de termos.

Uma representação temática, como é concebida nesta abordagem, é constituída por nodos temáticos. Cada nodo temático possui um termo selecionado como "centro temático" e outros termos (subtemas) semântica e tematicamente associados a este. A representação é construída através dos seguintes passos: identificação dos termos no texto, resolução de ambigüidades (através de um thesaurus de domínio), construção dos nodos temáticos e determinação do *status* dos nodos temáticos.

São considerados "centros temáticos" os descritores contidos no título e na primeira sentença do texto, bem como aqueles com elevada freqüência de ocorrência. Fazem parte de um mesmo nodo temático os descritores relacionados aos centros temáticos.

Para determinar o *status* de cada nodo temático, assume-se que os descritores dos nodos principais devem ser encontrados ao longo de todo o texto. São definidos nodos temáticos "principais" e "específicos", além de "conceitos mencionados".

Nodos temáticos principais são aqueles que (i) têm relações textuais com a maioria dos outros principais; e (ii) têm um somatório de freqüência dos descritores, relacionados entre os principais, maior do que o somatório de freqüência correspondente ao mesmo número de outros nodos temáticos do texto. Nodos temáticos específicos são

aqueles que incluem descritores presentes em, pelo menos, dois diferentes nodos temáticos principais. Finalmente, conceitos mencionados são descritores não presentes em nodos temáticos principais ou específicos.

Os descritores de um texto são subdivididos em classes, cada uma com seu peso: descritores principais de nodos temáticos principais (peso=1,00), outros descritores de nodos temáticos principais (peso=0,60), descritores principais de nodos temáticos específicos (peso=0,30), outros descritores de nodos temáticos específicos (peso=0,10) e descritores definidos como conceitos mencionados (peso=0,05). Os descritores que não se encaixam nessas classes recebem peso=0,001.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Loukachevitch, Dobrov e co-autores, deste trabalho correlato com nodos temáticos (TCNT), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos e relacionamentos como descritores, em TCNT são usados apenas termos.
- Em TCNT é usada lematização, como normalização morfológica, enquanto no modelo TR+ é usada nominalização.
- Em TCNT é implementada normalização léxico-semântica através de thesaurus de domínio, enquanto no modelo TR+ isso é feito através de autômato finito somente aplicável a alguns substantivos.
- Enquanto no modelo TR+ o peso dos descritores é calculado com base no conceito de evidência, em TCNT os pesos são predeterminados em função da classe do descritor.

2.6.4 Índice estruturado em árvore binária

Matsumura, Takasu e Adachi [MAT 2000] constroem um índice estruturado, representado por uma árvore binária, para recuperação de textos em língua Japonesa. Essas estruturas são construídas em três etapas: análise morfológica, análise de dependência baseada em padrões e análise de substantivos compostos.

Um relacionamento de dependência é constituído de uma *relation word*, que identifica o relacionamento, e duas *concept words*, que são os argumentos da relação. *Concept words* incluem substantivos, adjetivos, advérbios e constituintes de substantivos compostos. *Relation words* incluem preposições, verbos auxiliares e principais e suas combinações.

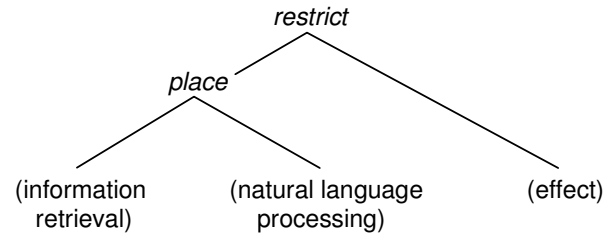

Figura 2.5: Exemplo de árvore binária

Relation words e *concept words* são extraídas do título e do resumo dos textos, no caso dos documentos, ou diretamente da consulta, expressa na forma de sintagma nominal. A análise de dependência, realizada manualmente, se baseia em padrões. Por exemplo: o padrão "c1 *of* c2 *on* c3" é substituído por "c1 *restrict* c2 *place* c3", onde c1,

c2 e c3 são *concept words*, e *restrict* e *place* representam *relation words*. A Figura 2.5 [MAT 2000] apresenta, como exemplo, uma árvore binária gerada a partir da sentença em Inglês "effect of natural language processing on information retrieval".

A análise dos substantivos compostos possibilita a especificação de relacionamentos entre os conceitos contidos nesses substantivos. É adotado o princípio que estabelece que os substantivos compostos podem ser transformados através da ligação de seus componentes por palavras funcionais. Por exemplo: em Inglês "*information retrieval*" pode ser transformado em "*retrieval of information*" e a análise de dependência pode ser novamente aplicada.

Considerando a parte *b* de um documento, constituída pelo seu título ou resumo, o valor de relevância de *b* (VR_b) é dado por:

$$VR_b = xw\sum_t \left(\log(f_{t,b}+1)IDF_t\right) + (1-xw)\sum_r \max\left(level_r\, IDF_{t1}\, IDF_{t2}\, gw_{t1}\, gw_{t2}\right) \qquad (3)$$

onde:

xw é um parâmetro entre 0 e 1;

t é um dos termos da parte b;

$f_{t,b}$ é a freqüência de ocorrência de t em b;

IDF_t é o fator *IDF* do termo t (o mesmo vale para t_1 e t_2);

r é um dos relacionamentos da parte b;

t_1 e t_2 são as *concept words* componentes de r;

a função max seleciona o maior peso de um relacionamento entre suas distintas ocorrências na parte b;

$level_r$ depende (conforme é explicado adiante) do nível de coincidência encontrado na comparação dos relacionamentos (da consulta e do documento); e

$gw_t = 0$, se t é um termo geral (como "effect", "action" ou "cause", em Inglês), ou 1, em caso contrário.

O nível correspondente ao fator $level_r$ pode ser *exact* (se as *relation words* são iguais), *category* (se as *relation words* são diferentes mas pertencentes à mesma categoria) ou *wild* (se as categorias e as *relation words* são diferentes).

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Matsumura, Takasu e Adachi, deste trabalho correlato com árvore binária (TCAB), podem ser destacados os seguintes aspectos:

- Ambas as estratégias usam termos e relacionamentos como descritores.
- No modelo TR+ são usadas regras gramaticais para a identificação de relacionamentos, incluindo normalização sintática. Em TCAB, por outro lado, é realizada normalização sintática através da transformação de substantivos compostos, após serem identificados.
- Ambas as estratégias assumem que o peso dos relacionamentos depende do peso dos termos, entretanto no modelo TR+ o peso dos termos também depende da quantidade de relacionamentos em que se envolvem.
- O modelo TR+ não rotula os identificadores de relacionamentos, ao contrário do que em TCAB é feito com as *relation words*, que substituem as proposições, por exemplo.
- Ao contrário do modelo TR+, em TCAB os procedimentos não são totalmente automatizados e é considerada somente parte (título e resumo) dos documentos.

2.6.5 Triplas com relações semânticas

Litkowski [LIT 2000] propõe a representação de textos através de triplas com relações semânticas, para melhorar a eficiência de sistemas questão-resposta[11]. A extração das triplas é realizada através dos seguintes passos: identificação das sentenças do texto, análise sintática de cada sentença e análise da árvore sintática para a geração das triplas.

O analisador sintático utilizado adota uma gramática com 350 regras e um dicionário com as categorias gramaticais das palavras.

As triplas, extraídas das árvores sintáticas construídas, são constituídas por uma entidade do discurso, uma relação semântica e uma palavra governante (*governing word*).

Entidades do discurso podem ser de diversas naturezas, como números, seqüências de adjetivos, pronomes possessivos e seqüências de substantivos.

As relações semânticas identificam os papéis semânticos das entidades. Estes papéis são, por exemplo, agente, tema, tempo e modificador. Tais papéis são identificados por termos como SUBJ, OBJ, TIME e ADJMOD. Também são usadas, como identificadores de relações, as preposições que encabeçam os sintagmas preposicionais.

As palavras governantes são, geralmente, aquelas com as quais as entidades do discurso se relacionam na sentença. No caso de SUBJ, OBJ e TIME, as palavras governantes são os verbos principais das sentenças. No caso de preposições, elas são os substantivos ou os verbos que os sintagmas preposicionais modificam. No caso de ADJMOD, elas são geralmente os substantivos modificados.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Litkowski, deste trabalho correlato com triplas (TCTR), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos e relacionamentos como descritores, em TCTR são usados apenas relacionamentos.
- Ambas as estratégias usam regras gramaticais para a identificação de relacionamentos, incluindo normalização sintática.
- Enquanto em TCTR o cálculo do peso dos descritores está baseado na freqüência de ocorrência, no modelo TR+ tal cálculo baseia-se no conceito de evidência.
- Ambas as estratégias usam preposições como identificadores de relacionamentos mas, ao contrário do que ocorre em TCTR, o modelo TR+ não considera relacionamentos rotulados com termos como SUBJ e OBJ.

2.6.6 Pares de termos lematizados

Arampatzis e co-autores [ARA 97, ARA 2000, ARA 2000a] avaliam diversas abordagens de indexação incluindo termos, bigramas (substantivo-adjetivo) e pares modificado-modificador. O texto é processado através de (i) identificação de sentenças e palavras, (ii) etiquetagem morfo-sintática, (iii) extração de sintagmas nominais, (iv) eliminação de *stopwords*, (v) transformação de sintagmas nominais de três ou mais palavras para duas palavras, e (vi) normalização morfológica.

---

[11] Do Inglês *question-answering*.

A transformação de sintagmas nominais se baseia na freqüência de ocorrência, criando pares de palavras com associação mais freqüente estatisticamente.

Duas abordagens para normalização morfológica foram testadas: lematização e *stemming*.

A diferença entre bigramas e pares modificado-modificador, nesta estratégia, pode ser explicada com o seguinte exemplo. Das frases em Inglês "air pollution" e "pollution of the air" seriam extraídos dois bigramas – "(air,pollution)" e "(pollution,air)" –, mas apenas um par modificado-modificador resultante de normalização sintática – "(pollution,air)".

O cálculo do peso de um descritor *i* em um documento *d* é estabelecido, no modelo vetorial, através da Equação 1, sendo $w_{i,d} = 0$, se $f_{i,d} = 0$, ou, caso contrário,

$$w_{i,d} = \log(f_{i,d}) + 1 \quad (4)$$

onde:

$f_{i,d}$ é a freqüência de ocorrência de *i* em *d*.

Levando em conta os resultados obtidos, o processo de lematização se mostrou, segundo Arampatzis e co-autores, mais eficiente que o de *stemming*. Relacionamentos na forma de bigramas tiveram melhores resultados que pares modificado-modificador com consultas curtas; o contrário aconteceu com consultas longas.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Arampatzis e co-autores, deste trabalho correlato com pares lematizados (TCPL), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos nominalizados e relacionamentos como descritores, em TCPL são usados pares de termos lematizados.
- Em TCPL é usada lematização (e *stemming,* como alternativa), como normalização morfológica, enquanto no modelo TR+ é usada nominalização.
- No modelo TR+ são usadas regras gramaticais para a identificação de relacionamentos, incluindo normalização sintática. Por outro lado, em TCPL a normalização sintática é apoiada por estatística, ocorrendo após a identificação dos relacionamentos.
- Enquanto em TCPL o cálculo do peso dos descritores adota esquema TF.IDF, no modelo TR+ tal cálculo se baseia no conceito de evidência.

2.6.7 Expressões ternárias

Katz e Lin [KAT 2000, LIN 2001] usam o sistema REXTOR (*Relations EXtracTOR*) para extrair expressões ternárias de textos, aplicáveis em sistema questão-resposta. São utilizadas duas classes de regras para extrair e compor essas expressões: regras de extração e regras de relação. As primeiras são utilizadas para identificar padrões no texto, de acordo com uma gramática. Essas regras descrevem padrões para entidades específicas como sintagmas preposicionais e grupos de substantivos. As regras de relação são ativadas pelas extrações bem sucedidas de entidades específicas. Uma gramática de relações guia a construção de cada expressão ternária.

As expressões ternárias podem ser vistas, de forma intuitiva, como triplas sujeito-relação-objeto. Elas podem ser expressas através de diversos tipos de relações, como relações sujeito-verbo-objeto, relações de posse ou outras. Do ponto de vista

sintático, essas expressões podem ser consideradas relações binárias. Do ponto de vista semântico, elas podem ser consideradas predicados com dois argumentos e, assim, podem ser manipuladas, segundo Katz e Lin, através da lógica de predicados.

Na Tabela 2.2 [KAT 2000, LIN 2001] são exemplificadas algumas expressões ternárias extraídas de sentenças e frases em Inglês.

Tabela 2.2: Exemplos de expressões ternárias

texto	expressões ternárias
shiny happy people of Wonderland	<shiny *describes* people> <happy *describes* people> <people *related-to* Wonderland>
the president surprised the country with his actions	<president *is-subject-of* surprise> <country *is-direct-object-of* surprise > <surprise *with* actions>
the meaning of life	<meaning *possessive-relation* life>
the bank near the river	<bank *near-relation* river>

Podem ser utilizadas regras transformacionais, aplicadas sobre expressões ternárias, para realizar normalização lingüística. Por exemplo, as expressões, em Inglês, <who shock country>, <shock with declaration> e <declaration related-to who> podem ser transformadas em <declaration shock country> e <declaration related-to who>. Assim, as sentenças "Whose declaration of guilt shocked the country?" e "Who shocked the country with his declaration of guilt?" seriam representadas através das mesmas expressões ternárias.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Katz e Lin, deste trabalho correlato com expressões ternárias (TCET), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos e relacionamentos como descritores, em TCET são usadas expressões ternárias.
- No modelo TR+ são usadas regras gramaticais para a identificação de relacionamentos, incluindo normalização sintática. TCET, entretanto, realiza normalização sintática através de regras transformacionais, após a identificação dos relacionamentos.
- Enquanto em TCET o cálculo do peso dos descritores está baseado na freqüência de ocorrência, no modelo TR+ tal cálculo se baseia no conceito de evidência.
- No modelo TR+ as preposições são identificadores predominantes dos relacionamentos e não há rótulos para as relações, como em TCET.

2.6.8 Famílias morfológicas e pares de dependência

Barcala, Vilares e co-autores [BAR 2002, VIL 2002] usam duas formas para reduzir a variedade lingüística: o uso de morfologia derivacional produtiva e a identificação de pares de termos com dependência sintática. Eles usam o conceito de família morfológica, definida como um conjunto de palavras com mesma raiz. São usados mecanismos de derivação, como derivação de morfemas, identificação de variantes alomórficas (que não apresentam similaridade morfológica), e derivação influenciada por condições fonológicas. Um léxico é utilizado para conferir a existência

de cada termo derivado. Cada família morfológica construída possui um termo representante que é utilizado no processo de normalização lexical.

Cada termo é normalizado tomando como base a família morfológica à qual pertence. Se uma palavra *p* pertence a uma família morfológica cujo representante é o termo *t*, então *t* será a forma normalizada adotada para *p*.

Os pares de termos com dependência sintática são extraídos com auxílio de gramática. Esses pares podem ser do tipo modificado-modificador, sujeito-verbo ou verbo-complemento, conforme os exemplos em Espanhol "(casa,viejo)", "(perro,comer)" e "(recortar,gasto)", respectivamente.

As expressões em Espanhol "recorte de gastos" e "recortar gastos", por exemplo, são normalizadas levando em conta os representantes das famílias morfológicas. Assim "(recorte,gastos)" e "(recortar,gastos)" são normalizadas como "(recorte,gastar)". Neste exemplo, fica subentendido que "recorte" é o representante da família de "recorte" e "recortar", enquanto que "gastar" é o representante da família de "gastos".

É utilizado o modelo vetorial com esquema TF.IDF tanto para termos quanto para relacionamentos, ou seja, os pares de termos com dependência sintática.

Barcala, Vilares e co-autores relatam haver testado alternativas para normalização lexical, como *stemming*, lematização e o processo proposto, com representante de família morfológica. Foram construídos índices com (i) palavras sem normalização, (ii) *stems*, (iii) lemas, (iv) pares de termos lematizados, e (v) pares de termos com representantes de famílias morfológicas. Os resultados, principalmente quanto à precisão, melhoram nesta ordem (i a v). Eles concluem que sua proposta favorece a performance de sistemas de RI, principalmente no caso de idiomas morfologicamente ricos como o Espanhol.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Barcala, Vilares e co-autores, deste trabalho correlato com famílias morfológicas (TCFM), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos e relacionamentos como descritores, na melhor alternativa em TCFM são usados apenas relacionamentos constituídos por pares modificado-modificador.
- TCFM usa, como normalização morfológica, o procedimento onde o termo corrente é substituído pelo representante da correspondente família morfológica. Por outro lado, no modelo TR+ é usada nominalização.
- Ambas as estratégias usam regras gramaticais para a identificação de relacionamentos, incluindo normalização sintática.
- Enquanto em TCFM o cálculo do peso dos descritores está baseado na freqüência de ocorrência, no modelo TR+ tal cálculo se baseia no conceito de evidência.

2.6.9 Bitermos

Srikanth e Srihari [SRI 2002] usam o conceito de bitermo para representar dependência de termos na abordagem probabilística de modelagem de linguagem. De acordo com esses autores, um bitermo é similar a um bigrama exceto pelo fato de a restrição de ordem dos termos ser atenuada, ou seja, bitermos são pares de termos não ordenados. Os relacionamentos são constituídos, então, por bitermos, sendo que um bitermo $\{i,j\}$ corresponde aos bigramas (i,j) e (j,i).

Dos esquemas que os autores desta estratégia testaram, o que é representado pela

Equação 5 produziu melhores resultados de recuperação. O peso de um bitermo $\{i,j\}$ em um documento d, nesse caso, é dado por:

$$W_{\{i,j\},d} = \alpha_1 \frac{f_{(i,j),d} + f_{(j,i),d}}{2\min\{f_{i,d}, f_{j,d}\}} + (1-\alpha_1)\left(\alpha_2 p_{j,d} + (1-\alpha_2) p_{j,C}\right) \quad (5)$$

onde:

$f_{i,d}$ é a freqüência de ocorrência do termo i em d (o mesmo vale para o termo j);

$f_{(i,j),d}$ é a freqüência de ocorrência de i quando antecedido por j em d (o mesmo vale para (j,i));

$\min\{f_{i,d},f_{j,d}\}$ retorna o menor valor entre $f_{i,d}$ e $f_{j,d}$;

$p_{j,d} = \frac{f_{j,d}}{td}$; $p_{j,C} = \frac{f_{j,d}}{tC}$;

td é o total de termos de d;

tC é o total de termos na coleção de documentos; e

α_1 e α_2 são parâmetros cujos valores são 0,1 e 0,4 respectivamente, no experimento de Srikanth e Srihari.

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Srikanth e Srihari, deste trabalho correlato com bitermos (TCBT), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos e relacionamentos como descritores, em TCBT são usados bitermos, ou seja, relacionamentos.
- Enquanto em TCBT o peso de um bitermo depende da ocorrência dos termos e dos bigramas, no modelo TR+ o peso dos relacionamentos depende do peso dos termos e, por sua vez, o peso dos termos depende da quantidade de relacionamentos em que se envolvem.

2.6.10 Conexões gramaticais

Changki Lee e Gary Lee [LEE 2005] adaptaram uma estratégia desenvolvida por Rijsbergen [RIJ 79] com a inclusão de conexões gramaticais através de árvore de análise sintática com dependências. Rijsbergen adotou o algoritmo proposto por Chow e Liu [CHO 68] para incorporar dependências de termos ao modelo probabilístico. Na abordagem de Rijsbergen é utilizada uma árvore geradora máxima baseada em medida de informação mútua esperada, considerando a distribuição de co-ocorrência dos termos na coleção. Com o uso de árvore de análise sintática, na estratégia de Changki Lee e Gary Lee, é reduzido o número de conexões necessárias e é otimizada a implementação do modelo. Na árvore, o relacionamento entre dois nodos (pai e filho) determina que o nodo filho é dependente (ou modificador) do nodo pai. Na frase, em Inglês, "brown dog", por exemplo, "dog" é o nodo pai e "brown" o nodo filho.

São utilizados somente termos como descritores de conceitos. Os relacionamentos servem apenas para determinar as dependências dos termos. Os pares modificado-modificador são extraídos dos textos e armazenados em uma base de dados. Essa base é pesquisada para verificar se há dependência, ao ser calculado o peso de cada termo, conforme o esquema descrito a seguir.

Para calcular a representatividade de um descritor i, o fator IDF_i, na Equação 2, é substituído por $dep_{i,d}$, calculado conforme a Equação 6. Considerando que a dependência entre os termos i (nodo filho) e j (nodo pai) é encontrada na consulta, $dep_{i,d}$ para o descritor i no documento d é dado por,

$$dep_{i,d} = x_i \log \frac{1}{q_i} + k_7 x_j \log \frac{q_j - q_{ij}}{q_j(1-q_i)} + k_8 x_i x_j \log \frac{q_i q_j}{q_{ij}} \quad (6)$$

onde:

$x_i = 1$ se i ocorre no documento d, 0 em caso contrário (o mesmo vale para x_j);

$q_i = \frac{df_i}{N}$; $q_j = \frac{df_j}{N}$; $q_{ij} = \frac{df_{ij}}{N}$;

df_i é o número de documentos onde i ocorre (o mesmo vale para df_j); e

df_{ij} é o número de documentos onde i e j ocorrem como par modificado-modificador, sendo j o modificado e i o modificador; e

k_7 e k_8 são parâmetros.

Embora o custo da recuperação aumente devido ao uso da árvore de análise sintática, tanto em fase de indexação (sobre os documentos), quanto em fase de busca (aplicada sobre a consulta), a performance do sistema melhora em relação ao simples uso da fórmula Okapi BM25 (Equação 2).

Ao ser comparada a estratégia do modelo TR+, proposto no próximo Capítulo, com a estratégia de Changki Lee e Gary Lee, deste trabalho correlato com pares modificado-modificador (TCMM), podem ser destacados os seguintes aspectos:

- Enquanto no modelo TR+ são usados termos e relacionamentos como descritores, em TCMM são usados apenas termos. Os relacionamentos têm utilidade somente no cálculo do peso dos termos.
- Em ambas as estratégias o peso dos termos depende dos relacionamentos, entretanto no modelo TR+ o peso dos relacionamentos também é calculado, porque são considerados descritores, ao contrário do que ocorre em TCMM, e depende do peso dos termos.

2.6.11 Análise comparativa dos trabalhos correlatos

A seguir são apresentados na Tabela 2.3, de forma integrada, dados sobre cada trabalho correlato descrito e, também, sobre a presente proposta, quanto aos descritores utilizados e ao esquema adotado para o cálculo da representatividade de cada descritor. Na Tabela 2.4, o mesmo é feito quanto ao modelo e as normalizações lexical e sintática. Alguns trabalhos descritos apresentam alternativas e, nesses casos, a alternativa de melhor performance é considerada.

Esta análise comparativa destaca os trabalhos correlatos. Ela será retomada na Seção 8.2, sob o enfoque do modelo proposto, após a apresentação e a avaliação do mesmo.

Na Tabela 2.3, a coluna “descritores” contém informações sobre os tipos de descritores utilizados em cada estratégia. Na coluna “peso dos descritores” são identificados os principais fatores que influenciam o cálculo da representatividade dos descritores.

Dentre os trabalhos descritos, somente TCIM (Seção 2.6.2), com índices múltiplos, menciona a utilização de arquivos de índice separados para buscas independentes. Este também é o caso do modelo TR+. TCMM (Seção 2.6.10), com pares modificado-modificador, usa um arquivo de índices, para os termos, e uma base de dados com informações para relacionar termos modificados e modificadores.

Assim como no modelo TR+, apenas TCIM (Seção 2.6.2), com índices múltiplos, e TCAB (Seção 2.6.4), com árvore binária, adotam tanto termos quanto

relacionamentos como descritores, e especificam cálculo de peso para ambos.

Tabela 2.3: Análise comparativa dos trabalhos correlatos (I)

estratégia	descritores	peso dos descritores
TCEI	termo-conector-termo	não mencionado
TCIM	termos, termos compostos, nomes próprios e pares modificado-modificador	esquema TF.IDF com incremento de peso para os relacionamentos
TCNT	descritores principais (ou não) de nodos (temáticos e específicos) além de conceitos mencionados	pesos por classe de descritor
TCAB	termos e relacionamentos formados por duas *concept words* e uma *relation word*	peso dos relacionamentos depende do peso dos termos
TCTR	triplas com relações semânticas	freqüência de ocorrência
TCPL	pares de termos lematizados	esquema TF.IDF
TCET	expressões ternárias	freqüência de ocorrência
TCFM	pares modificado-modificador	freqüência de ocorrência
TCBT	bitermos	peso do bitermo depende da ocorrência dos termos e dos bigramas
TCMM	termos	peso dos termos depende das conexões gramaticais
modelo TR+	termos nominalizados e relações lexicais binárias	evidência dos termos depende dos relacionamentos e vice-versa

Na Tabela 2.4, a coluna "modelo" identifica o modelo de espaço de descritores adotado pelas estratégias. A coluna "normalização lexical" contém informações sobre o uso de processos de conflação ou procedimentos alternativos para normalização lexical. Na coluna "normalização sintática" aparecem indicações sobre a abordagem utilizada para a normalização dos relacionamentos.

Tabela 2.4: Análise comparativa dos trabalhos correlatos (II)

estratégia	modelo	normalização lexical	normalização sintática
TCEI	RT	sem informação	não
TCIM	TR	semelhante a *stemming*	regras gramaticais
TCNT	TT	lematização e sinonímia	não
TCAB	TR	lematização*	transformação de substantivos compostos
TCTR	TR	lematização*	regras gramaticais
TCPL	RT	lematização	apoiada por estatística
TCET	RT	lematização*	regras tranformacionais
TCFM	TR	com representante de família morfológica	regras gramaticais
TCBT	NG	sem informação	não
TCMM	TT	lematização*	não
modelo TR+	TR+(RT)	nominalização	regras de identificação das relações lexicais binárias

* deduzido dos exemplos apresentados pelos autores.

Somente nas estratégias de TCNT (Seção 2.6.3), com nodos temáticos, e do modelo TR+ há normalização léxico-semântica. TCNT usa sinonímia através de um thesaurus de domínio. Os outros trabalhos executam somente normalização morfológica

e usam lematização ou *stemming*. TCIM (Seção 2.6.2), com índices múltiplos, inclui relações semânticas através de expansão de consulta.

Dos dez trabalhos descritos, quatro não realizam normalização sintática. Além do modelo TR+, apenas os trabalhos que utilizam regras gramaticais produzem os relacionamentos já normalizados sintaticamente. Os restantes – TCPL (Seção 2.6.6), com pares lematizados, TCET (Seção 2.6.7), com expressões ternárias, e TCAB (Seção 2.6.4), com árvore binária – normalizam os relacionamentos após serem identificados.

O modelo TR+, proposto no próximo Capítulo, combina o modelo TR com aspectos do modelo RT, quanto à influência entre termos e relacionamentos no cálculo da representatividade desses descritores. Esta é uma das diferenças mais importantes entre a presente proposta e os trabalhos correlatos descritos. Esta diferença existe em razão das relações lexicais binárias e do conceito de evidência.

2.7 Resumo do Capítulo

Um espaço de descritores é definido como um conjunto de termos e, em alguns casos, um conjunto de relacionamentos, tendo cada um desses descritores um grau de representatividade quanto ao texto descrito. Enquanto termos descrevem conceitos atômicos, relacionamentos descrevem conceitos formados por outros conceitos. No caso do espaço de descritores do modelo com unigramas, o conjunto de relacionamentos é vazio.

Em fase de indexação, após a *toquenização*, acontece a seleção dos descritores, onde é usual a eliminação de s*topwords*. Processos de normalização lexical e sintática podem ser incluídos durante ou após a seleção dos descritores. Na normalização lexical (morfológica ou léxico-semântica), descritores são gerados com maior expressividade ao serem eliminadas variações morfológicas e reconhecidas famílias de palavras com mesmo significado. Na normalização sintática, há a transformação de frases semanticamente equivalentes mas sintaticamente diferentes, em formas únicas e representativas.

Na normalização morfológica, a redução das formas flexionais a um só termo é usual através de conflação, geralmente através dos processos de *stemming* ou lematização. Esses processos reduzem o número de descritores e o tamanho do arquivo de índice, e tornam a recuperação independente da forma com que o termo ocorre na consulta. Entretanto, alguns problemas persistem: certos significados podem ser perdidos no *stemming* e palavras de famílias de significados diferentes podem ser agrupadas. Enquanto um algoritmo simples de *stemming* para Inglês é suficiente, estratégias mais sofisticadas são necessárias para idiomas com morfologia flexional mais complexa, como o Português. Na lematização, por outro lado, algumas palavras pertencentes à mesma família de significados podem não ser normalizadas.

O grau de representatividade de um descritor é calculado através de fórmulas específicas para as abordagens vetorial e probabilística, baseadas em freqüência de ocorrência, com penalização para a verbosidade dos documentos e para descritores muito freqüentes na coleção.

Diversos pesquisadores têm analisado a aplicação de abordagens estatísticas e de conhecimento lingüístico na representação de dependências de termos. As estratégias desenvolvidas que não adotam o modelo UG (Figura 2.2), com unigramas, seguem quatro modelos alternativos:

- modelo NG, com representação de texto concebida como um conjunto de

relacionamentos estatísticos, onde a representatividade dos relacionamentos depende da co-ocorrências dos termos;

- modelo TT, com representação de texto concebida como um conjunto de termos relacionados estatística ou sintaticamente, onde a representatividade dos termos depende da sua freqüência de ocorrência e do termo relacionado;
- modelo TR, com representação de texto concebida como um conjunto de termos relacionados sintaticamente, principalmente, através de ligações modificado-modificador, onde a representatividade dos relacionamentos depende dos termos; e
- modelo RT, com representação de texto concebida como um conjunto de relacionamentos sintáticos e seus componentes, onde a representatividade dos termos depende dos relacionamentos. Sintagmas nominais constituem os principais descritores neste caso.

São apresentados quatro trabalhos correlatos que seguem o modelo TR (com índices múltiplos, com árvore binária, com triplas e com famílias morfológicas), três que seguem o modelo RT (com expressões de índice, com pares lematizados e com expressões ternárias), dois que seguem o modelo TT (com nodos temáticos e com conexões gramaticais) e um que segue o modelo NG (com bitermos).

No próximo Capítulo é proposto um novo modelo com dependência de termos para RI.

3 PROPOSTA: TERMOS E RELACIONAMENTOS EM EVIDÊNCIA

3.1 Introdução

Na teoria da informação encontramos subsídios para entender os indicadores da quantidade de informação [MAC 73]. A informação (ou a variação) de um conjunto de possibilidades é medida pelo logaritmo binário do número de possibilidades [SHA 47]. Salton e MacGill [SAL 83] já afirmavam que o conteúdo de informação de uma palavra é medido por $-log_2 p$, onde p é a probabilidade de ocorrência da palavra.

A estatística tem sido utilizada para o cálculo da representatividade dos descritores, levando em conta a freqüência de ocorrência dos mesmos. Segundo Luhn, apud [RIJ 79], "a freqüência de ocorrência de uma palavra no texto fornece uma medida útil de sua significância". Entretanto, Swanson [SWA 88] já advertiu que "dados estatísticos sobre ocorrência de palavras não representam nem podem ser substituídos por significado. Podem, porém, com sucesso ocasional sinalizar ou indicar áreas potencialmente frutíferas onde um ser humano pode buscar significado ou relevância".

O que há além da freqüência de ocorrência das palavras? Um texto não pode ser concebido apenas como uma seqüência de caracteres ou palavras. Diversos pesquisadores têm estudado as relações que acontecem entre os componentes de um texto. Por exemplo, relações essenciais de significado de Porzig (apud [LYO 77]) são descritas como relações binárias formadas por pares de lexemas conectados sintagmaticamente. Neste mesmo sentido, as colocações (relações sintagmáticas) são relações que codificam informações sobre os componentes lexicais [JUR 2000]. Segundo Halliday e Hasan [HAL 76], "um texto tem textura e é isto que o distingue de um não-texto".

Como capturar a textura de um texto? "A textura (de um texto) é produzida por relações coesivas. Qualquer segmento de um texto pode ser caracterizado através do número e do tipo das relações coesivas que ele apresenta" [HAL 76]. Embora as relações coesivas sejam alvo de análise em diversos trabalhos [FÁV 97, LOU 99], será adotada, neste trabalho, a classificação que identifica dois grandes grupos de relações coesivas [MIR 2003]: gramaticais e lexicais.

São enfatizadas, aqui, as relações gramaticais identificadas entre os mecanismos de coesão frásica [MIR 2003] ao serem analisados os termos essenciais (sujeito e predicado), integrantes (complementos nominais e verbais e agente da passiva) e acessórios (aposto e adjuntos adnominais e adverbiais) das orações [CEG 91, SAC 99].

Os mecanismos de coesão frásica constituem a fonte para a identificação das relações lexicais binárias utilizadas nesta proposta. Essas relações usam freqüentemente, como elemento de ligação, as preposições que Aristóteles já destacava como categoria morfológica importante para esta função ao descrever os "relativos": "São chamados 'relativos', então, os termos cuja natureza se explica com referência a algo mais, a preposição 'de' ou outra preposição sendo usada para indicar esta relação" [MAC 73]. As preposições têm papel fundamental no mapeamento de dependências sintáticas [GAM 2002], sendo importantes identificadores de relações lexicais binárias.

Esta tese sugere então que, além da freqüência de ocorrência, as palavras têm determinadas características morfológicas e relações de coesão úteis para o cálculo da representatividade dos descritores. O modelo para RI aqui proposto, com dependência de termos e denominado TR+, adota o conceito de evidência para capturar essas características. O modelo TR é combinado com aspectos do modelo RT, pois a representatividade dos termos, em TR+, afeta a representatividade dos relacionamentos e vice-versa. A evidência de um termo depende de sua participação em mecanismos de coesão frásica, e a evidência de um relacionamento depende das evidências de seus argumentos (termos).

Termos e relacionamentos, portanto, são considerados descritores, mas quais termos? Qual o critério de seleção? Qual a alternativa para os processos usuais de conflação? Pela teoria clássica dos nomes de Mill (apud [CAM 2004]), os nomes (os substantivos) são termos categoremáticos, ou seja, têm significação por si mesmos. Eles possuem a propriedade de denotar ou designar entidades, qualidades, estados, situações e ações, coisas reais ou imaginárias. Os substantivos, portanto, têm as qualidades típicas de um descritor.

Substantivos são, em geral, as palavras mais representativas do conteúdo de um texto [ZIV 99]. Confirmando a importância dos substantivos como descritores de conceitos, Lapata [LAP 2002] acredita que o reconhecimento da vinculação semântica que existe entre verbos e substantivos permite melhorar a classificação de documentos na RI. Lapata exemplifica que, com esta visão, é possível recuperar documentos contendo "to treat cancer" a partir de consultas como "cancer treatment", em Inglês.

Um processo de normalização morfológica que deriva substantivos de palavras de outras categorias morfológicas pode ser utilizado como alternativa para *stemming* e lematização. Este é o processo de nominalização.

Além das relações lexicais binárias, da nominalização e do conceito de evidência, o modelo proposto também prevê a inclusão de operadores Booleanos na consulta para complementar a especificação das dependências de termos.

3.2 Modelo TR+

O modelo TR+ não se restringe a um modelo de representação de texto: ele inclui, além da definição do espaço de descritores conforme segue, especificações para a formulação da consulta e para a classificação dos documentos recuperados. Portanto o modelo TR+ é um modelo de RI.

Definição: Espaço de descritores no modelo TR+

Dados uma coleção de documentos $D = \{d\}$, um conjunto de termos $T = \{t\}$ e um conjunto de relacionamentos $R = \{r\}$, um espaço de descritores I_{TR+} é definido como:

$$I_{TR+} = T \cup R$$

onde:

$T = \{t, t = \eta_1(p)$ ou $t = \eta_2(p)$ com peso $W_{t,d}$ em $d \in D\}$;
$R = \{r, r = id(t_1 \in T, t_2 \in T)$ com peso $W_{r,d}$ em $d \in D\}$;
p é uma palavra (adjetivo, particípio, verbo ou advérbio) contida em d;
η_1 e η_2 são operações de nominalização;
$id(t_1,t_2)$ é uma relação lexical binária (RLB); e
$W_{t,d}$ e $W_{r,d}$ são calculados com fórmula de cálculo baseada no conceito de evidência.

Na Figura 3.1 é apresentada uma visão geral do modelo TR+, com as etapas necessárias para gerar o espaço de descritores, em fase de indexação, e para a classificação por relevância dos documentos em relação à consulta, em fase de busca.

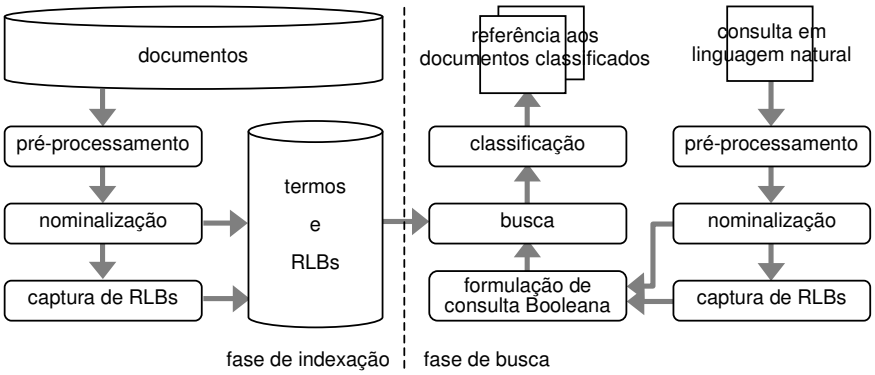

Figura 3.1: Visão geral do modelo TR+

No modelo TR+, os documentos e a consulta em linguagem natural recebem tratamento idêntico para a construção do espaço de descritores, no caso dos documentos, e para iniciar a formulação da consulta Booleana. Primeiramente, o texto é pré-processado com *toquenização* e etiquetagem morfológica, após é realizada a nominalização com a definição dos termos e, então, são identificadas as RLBs.

O espaço de descritores é construído com termos nominalizados e RLBs, sendo estas organizadas de acordo com o tipo para agilizar a pesquisa. O processamento da consulta fornece as mesmas informações para que sejam introduzidos os operadores Booleanos.

Durante a busca, termos e RLBs da consulta são pesquisados no espaço de descritores e é calculado o valor de relevância para os documentos de acordo com o peso dos descritores e os operadores Booleanos predefinidos.

Finalmente, os documentos são classificados por relevância decrescente em relação à consulta.

O processo de nominalização, as RLBs e o conceito de evidência para o cálculo da representatividade dos descritores, que caracterizam o modelo TR+, são discutidos a seguir. A proposta, entretanto, não se restringe a um modelo de espaço de descritores, ela especifica um modelo de RI. Assim, além dos aspectos que fazem parte da fase de indexação, são apresentados no restante deste Capítulo a formulação da consulta Booleana e o procedimento de classificação dos documentos recuperados, que fazem parte da fase de busca e complementam o modelo TR+.

3.3 Nominalização

Em sentido amplo, nominalização é um processo de formação de palavras onde um novo substantivo é derivado de uma palavra existente no léxico, principalmente verbos e adjetivos [KEH 2000]. No modelo TR+, nominalização é a transformação de uma palavra (adjetivo, particípio, verbo ou advérbio), existente no texto, em um substantivo semanticamente correspondente, formado através de regras válidas de formação de palavras. Assim, a derivação de “amigavelmente” em “amigo” é uma

nominalização válida, ainda que na língua Portuguesa, neste caso, o substantivo seja a palavra original e o advérbio seja a palavra derivada.

As operações de nominalização η_1 e η_2, conforme o modelo TR+, derivam respectivamente substantivos abstratos e concretos. Os primeiros representam eventos, qualidades, estados, ou outras entidades abstratas que podem ser derivadas de adjetivos, particípios, verbos ou advérbios, como:

η_1(encontrar) = encontro,

η_1(belo) = beleza e

η_1(livre) = liberdade.

Substantivos concretos, por outro lado, representam, geralmente, agentes derivados de verbos ou entidades derivadas de adjetivos, como:

η_2(pacificar) = pacificador e

η_2(continental) = continente.

Na Tabela 3.1 são apresentados alguns exemplos de nominalização. A partir de verbos (inclusive particípios), são comumente derivados substantivos abstratos e concretos. A partir de adjetivos e advérbios, as nominalizações são geralmente abstratas.

Tabela 3.1: Exemplos de nominalização

palavra original	classe	substantivo abstrato	substantivo concreto
saltar	verbo	salto	saltador
emendado	particípio	emenda	emendador
puro	adjetivo	pureza	ε
facilmente	advérbio	facilidade	ε
oval	adjetivo	ε	ε
fluvial	adjetivo	ε	rio
jovem	adjetivo	juventude	jovem

ε = ausência de nominalização

A partir de algumas palavras, como o adjetivo "oval" (ver Tabela 3.1), não são derivados substantivos (nem abstratos, nem concretos). A comparação entre as palavras "oval" e "fluvial" ajuda a explicar essa impossibilidade de nominalização. Ao contrário da equivalência existente em "barco fluvial" e "barco de rio", o mesmo não ocorre em "barco oval" e "barco de ovo". Pela mesma razão "pastel oval" não é equivalente a "pastel de ovo". Isso acontece porque "oval" não está diretamente relacionado a "ovo", mas a seu formato ("em forma de ovo"). Neste caso, o adjetivo ("oval") é considerado um descritor (do conceito envolvido) mais adequado que o substantivo ("ovo").

Há palavras que são lexicalmente idênticas aos seus substantivos derivados. Um exemplo é a palavra "jovem" (ver Tabela 3.1), que pode ocorrer tanto na forma de adjetivo quanto na de substantivo.

Os termos incluídos no espaço de descritores, no modelo TR+, são substantivos (ou palavras de outras classes, na ausência de nominalização) originais do texto, normalizados por lematização, e substantivos derivados por nominalização de adjetivos, advérbios, particípios e verbos. Esses termos são aqui denominados, indistintamente, "termos nominalizados", por serem formados essencialmente por substantivos (ou com semelhante poder de descrição). Eles constituem os argumentos das RLBs descritas a seguir.

Antes, deve ser salientado que os termos nominalizados são concebidos sem acentos e em minúsculas. Não foram estudados, neste trabalho, os efeitos da ausência de acentuação. De qualquer forma, foi adotada a hipótese de que podem ser esperadas, neste sentido, vantagens (por exemplo, "fábrica" e "fabrica" estão associados a um só conceito: o de fabricação) e desvantagens (por exemplo, "pública" e "publica", embora com raiz comum, estão associados a conceitos diferentes: o primeiro relativo à condição de algo destinado ao povo ou à coletividade, e o segundo representando o ato de publicar, editar, divulgar).

3.4 Relações Lexicais Binárias

RLBs [GON 2005] são relacionamentos entre termos nominalizados que capturam mecanismos de coesão frásica. Uma RLB possui a forma:

$$id(t_1,t_2)$$

onde:

id é um identificador de relação e
t_1 e t_2 são argumentos (termos nominalizados).

Há três tipos de RLBs, quanto ao identificador *id*:

- Classificação: quando *id* é especificado através do sinal de igual (=), t_1 representa uma subclasse ou uma instância de t_2 e t_2 representa uma classe.
  Exemplos: =(cao,animal)
  =(rex,cao)
- Restrição: quando *id* é uma preposição, t_1 representa um elemento modificado e t_2 representa um elemento modificador.
  Exemplos: de(equipe,atletismo)
  com(supervisor,experiencia)
  de(eficiencia,sistema)
  por(orientacao,ministro)
- Associação: quando *id* representa um evento, t_1 é um sujeito e t_2 é um objeto (direto ou indireto) ou um adjunto.
  Exemplos: superacao(aluno,dificuldade)
  interesse.a(proposta,negociante)
  moradia.em(papa,roma)

Nos dois últimos exemplos apresentados é possível constatar que um *id* de RLB do tipo associação pode ser preposicionado. Esta é uma estratégia que mantém a relação com dois argumentos mesmo em casos como esses. Na verdade, pode-se admitir que as RLBs dos tipos restrição e associação apresentam *ids* com o formato geral

evento.preposição

onde *evento* (ou *preposição*) pode estar ausente. Note que a RLB do tipo restrição "de(equipe,atletismo)" poderia ser concebida no formato "é.de(equipe,atletismo)" ou, como acontece, simplesmente ser assumida a ausência do "evento" neste caso. De modo semelhante, na RLB do tipo associação "superacao(aluno,dificuldade)" pode ser assumida a ausência da preposição.

RLBs são utilizadas no cálculo do peso dos descritores (sem distinção do tipo) como indicação de presença de seus argumentos em mecanismos de coesão. Por outro lado, a identificação do tipo da RLB (classificação, restrição e associação) é útil para

organizá-las em arquivos de índice distintos com o objetivo de agilizar a pesquisa das mesmas em fase de busca na RI.

RLBs são relações assimétricas onde cada argumento tem papel específico. Elas formam uma estrutura de relacionamentos que representa o texto dos documentos (ver exemplos no Anexo D).

Uma RLB pode ser classificada, também, quanto à nominalização de seus componentes, como:

- Original, quando nenhum dos componentes sofreu nominalização.

  Exemplos: calma do local $\xrightarrow{rlb}$ de(calma,local)

  rapidez da saída $\xrightarrow{rlb}$ de(rapidez,saida)

  representação do ator $\xrightarrow{rlb}$ de(representação, ator)

  cantor Zeviola $\xrightarrow{rlb}$ =(zeviola,cantor)

- Derivada, quando pelo menos um dos seus componentes é resultado de nominalização.

  Exemplos: local calmo $\xrightarrow{rlb}$ de(calma,local)

  saiu rapidamente $\xrightarrow{rlb}$ de(rapidez,saida)

  ator representou $\xrightarrow{rlb}$ de(representação, ator)

  Zeviola cantou $\xrightarrow{rlb}$ =(zeviola,cantor)

As RLBs são incluídas no espaço de descritores para dar maior cobertura de descrição. RLBs descrevem relações semânticas como as presentes na estrutura Qualia da teoria do Léxico Gerativo de James Pustejovsky [PUS 95, GON 2004]. Essas relações motivam a especificação e o uso das RLBs por terem aplicação na RI [GON 2001, GON 2001a, GON 2001b]. A estrutura Qualia, na teoria do Léxico Gerativo, descreve um item lexical α através de quatro papéis: formal, constitutivo, agentivo e télico.

O papel formal distingue α em um amplo domínio. Em uma RLB do tipo classificação, como "=(computador,maquina)" por exemplo, o computador seria distinguido como uma máquina ou, em "=(ipmf,tributo)", o ipmf seria um tributo.

O papel constitutivo indica o que faz parte de α. Em uma RLB do tipo restrição, como "de(casa,pedra)" por exemplo, haveria a indicação de que a casa é feita de pedra ou, em "com(massa,alho)", de que há alho na massa.

O papel agentivo especifica qual a razão de α passar a existir. Em uma RLB do tipo restrição, como "por(publicacao,autor)" por exemplo, seria especificado que a publicação se deve ao autor ou, em "por(impedimento,lei)", que a lei é a razão do impedimento.

O papel télico explica qual a função ou propósito de α. Em uma RLB do tipo associação, como "conserto(encanador,vazamento)" por exemplo, estaria a explicação de que a função do encanador é o conserto do vazamento ou, em uma RLB do tipo restrição como "para(leitura,aprendizado)", que o propósito da leitura é o aprendizado.

Entretanto, não se pretende que as RLBs "interpretem" o texto com distinções, indicações, especificações ou explicações dos tipos apresentados acima. Há a intenção de que as RLBs sejam descritores de tais fatos, mas sem rótulos. Por esta razão os identificadores de relação não são rotulados com este ou aquele papel. A única exceção é o identificador das RLBs do tipo classificação. O indicador "=" é o rótulo inevitável para o clássico "é um" porque não há outro papel possível nesse tipo de relação.

Em outras palavras, não se pretende que, por exemplo, a preposição "para" indique propósito, ou que a preposição "por" represente agentividade. Elas estão presentes em RLBs porque foram capturadas assim através de regras de identificação (ver Anexo B) e descrevem algo que o sistema não interpreta, mas que, mesmo que representado em ocorrências sintáticas distintas, deve ser descrito da mesma forma. Pretende-se, portanto, que cada RLB produza uma descrição independente da forma sintática com que os conceitos descritos estejam representados no texto.

As RLBs fundamentam o conceito de evidência aplicado aos descritores, conforme é discutido a seguir.

3.5 Evidência

Evidência é a condição do que se destaca, é a qualidade do que é evidente e, por sua vez, evidente é aquilo que não oferece ou não dá margem à dúvida [FER 99, HOU 2002]. O conceito de evidência [GON 2005] constitui um dos aspectos essenciais desta tese. O modelo TR+ adota cálculo baseado em evidência para o peso dos descritores, e não puramente em freqüência de ocorrência. No modelo proposto, a representatividade de um descritor depende, além de sua freqüência de ocorrência no texto, da ocorrência de mecanismos de coesão frásica.

No modelo TR+, o grau de representatividade de um descritor é afetado (i) pelo processo de nominalização, (ii) pela capacidade das regras para identificação de RLBs de deduzir estruturas de dependência evidentes e (iii) pela formulação do cálculo do peso dos descritores.

Como em qualquer processo de normalização lexical, a nominalização afeta a representatividade dos descritores, pois reúne palavras diferentes na forma de um único descritor. Os descritores normalizados tendem a ter pesos maiores que os não normalizados, por representarem grupos de palavras e, assim, acumular a freqüência de ocorrência das mesmas. É um processo que pode ser incluído na clássica transformação de termos "pobres" em "bons" [SAL 83].

As regras para identificação de RLBs afetam a representatividade dos descritores porque têm capacidade de reconhecer apenas estruturas de dependência evidentes. São tratadas como "não evidentes" as dependências à direita preposicionadas, após a segunda preposição[12]. Estas não são identificadas. Considere os seguintes exemplos:

(a) "arrombamento do cofre com explosivos"

(b) "arrombamento do cofre com jóias"

Nestes exemplos é identificada somente a RLB "de(arrombamento,cofre)". As RLBs "com(arrombamento,explosivo)", "com(cofre,explosivo)", "com(arrombamento,joia)" e "com(cofre,joia)" não são reconhecidas. Desta forma, alguns descritores perdem peso: as próprias RLBs não identificadas e os termos nelas presentes como argumentos. Assim, os termos "explosivo" e "joia" são penalizados, pois não são considerados participantes de um mecanismo de coesão reconhecido, ou seja, evidente.

Esta decisão se justifica pelo próprio conceito de evidência: deve haver destaque e não pode haver dúvida. Para um leitor humano, em (b) pode não haver dúvida e, conforme o contexto, em (a) também não, mas não se trata de processamento humano.

[12] As regras de captura podem identificar dependências não preposicionadas além desse limite. Por exemplo, a RLB "de(inutilidade,explosivo)", capturada de "arrombamento do cofre com explosivos inutilizados", é encontrada após a segunda preposição ("com").

De qualquer forma, considera-se que há carência de destaque, ou seja, que é dada pouca evidência aos termos "joia", em (b), e "explosivo", em (a). Há duas vantagens decorrentes desta decisão: (i) diminuição do esforço computacional necessário, ao deixar de resolver tais ambigüidades, e (ii) atendimento ao conceito de evidência, ao influenciar o cálculo do peso dos descritores, fazendo com que, quanto maior a evidência (destaque sem ambigüidade), maior seja a representatividade.

Quanto ao cálculo do grau de representatividade dos descritores, considerando o exemplo (a), a RLB "de(arrombamento,cofre)" fica evidenciada por 3 descrições: as dos conceitos correspondentes a cada um de seus argumentos e a descrição do relacionamento entre eles, ou seja, há um arrombamento, há um cofre e o arrombamento foi do cofre. Portanto, esta RLB receberia 3 unidades de evidência. Cada ocorrência dos descritores "arrombamento" e "cofre" receberia 1½ unidade, que é metade do valor atribuído à RLB. Há um arrombamento e é dito o que foi arrombado e, também, há um cofre e houve algo com ele. Finalmente, o descritor menos evidente, "explosivo", assim como "joia" no exemplo (b), receberia ½ unidade de evidência, sendo penalizado em 1 unidade por falta de coesão evidente. Há explosivos, mas eles foram usados no arrombamento ou estavam no cofre? De qualquer forma, os explosivos (assim como as jóias) são coadjuvantes nesta "parte da história". Pode ser que o autor (antes ou depois, no texto) os torne mais evidentes (ou não), envolvendo-os em outros mecanismos de coesão. Se isto acontecer, tais relacionamentos poderão vir a ser considerados.

Cada ocorrência de nova coesão gera 1 unidade de evidência, que é acrescida ao descritor envolvido. A RLB "de(arrombamento,cofre)", se fosse identificada em "arrombamento violento do cofre", receberia 4 unidades de evidência. Neste caso, "arrombamento" receberia 2½ unidades e "cofre" continuaria com 1½. Considera-se que, neste caso, "arrombamento" apresenta maior evidência pois dele se diz mais (foi realizado no cofre e com violência) que no caso anterior, ou seja, o texto tem mais a oferecer a uma possível consulta sobre ele.

A soma das unidades de evidência dos termos constituintes é sempre igual ao total de unidades de evidência da RLB constituída, em cada ocorrência. Esta propriedade é utilizada pela consulta Booleana e mencionada na Seção 3.6. Em resumo, os termos t_1 e t_2 e a RLB r, encontrados em uma consulta q, têm dupla contribuição no cálculo do valor de relevância de um documento d, caso t_1 e t_2 estejam relacionados através de r em d. Do contrário, se t_1 e t_2 ocorrem em d mas não estão relacionados através de r, a contribuição é simples e, assim, d tende a perder posições na classificação por relevância a q.

3.5.1 Cálculo de peso dos descritores e do valor de relevância

É adotada a abordagem probabilística (expressa na Equação 2) para o cálculo do peso dos descritores neste trabalho por ser mais eficiente quanto aos resultados de RI. Entretanto, o modelo TR+ é compatível também com a abordagem vetorial (expressa na Equação 1).

A Equação 7, uma adaptação da Equação 2 com a retirada do fator *IDF*[13], é adotada pelo modelo TR+. O peso $W_{i,d}$ do descritor i no documento d é dado por:

[13] A inclusão do fator *IDF* não beneficiou os resultados experimentais obtidos.

$$W_{i,d} = \frac{w_{i,d}(k_1 + 1)}{k_1((1-b) + b\frac{DL_d}{AVDL}) + w_{i,d}} \quad (7)$$

onde:

$w_{i,d}$ é a evidência do descritor i no documento d;

k_1, b, DL_d e $AVDL$ são os mesmos componentes utilizados na fórmula Okapi BM25, apresentada na Equação 2 (página 22).

A evidência $w_{i,d}$, representada através de $w_{t,d}$ para um termo t em um documento d, é calculada da seguinte forma no modelo TR+:

$$w_{t,d} = \frac{f_{t,d}}{2} + \sum_r f_{r,t,d} \quad (8)$$

onde:

$f_{t,d}$ é a freqüência de ocorrência de t em d e

$f_{r,t,d}$ é a quantidade de RLBs onde t é argumento em d,

e para uma RLB r, a evidência $w_{i,d}$ em um documento d, representada por $w_{r,d}$, é:

$$w_{r,d} = f_{r,d}(w_{t1,d} + w_{t2,d}) \quad (9)$$

onde:

$f_{r,d}$ é a freqüência de ocorrência de r em d e

$w_{t,d}$ é a evidência do argumento t de r em d;

Exemplos apresentados no Anexo D mostram diferenças nos resultados do cálculo baseado em evidência, em comparação à formulação baseada apenas em freqüência de ocorrência.

Termos e RLBs de uma consulta q são obtidos e têm seus pesos calculados da mesma forma utilizada para os documentos. Entretanto, para cada RLB r encontrada na consulta q, sendo

$$r = id(t_1,t_2),$$

é incluída uma RLB r' na consulta Booleana qb, sendo

$$r' = id'(t_1,t_2),$$

onde id' é qualquer identificador diferente de id (conforme é exemplificado na Seção 3.6). O peso $W_{r',q}$ de r' depende do peso $W_{r,q}$ de r, sendo penalizado por possuir identificador diferente, embora r e r' tenham mesmos argumentos. $W_{r',q}$ é dado por:

$$W_{r',q} = \frac{W_{r,q}}{2} \quad (10)$$

O valor de relevância $VR_{d,q}$ de um documento d para uma consulta q é obtido, no modelo TR+, por:

$$VR_{d,q} = \sum_i \left(W_{i,d} W_{i,q}\right) \quad (11)$$

onde:

$W_{i,d}$ é o peso de termos e/ou RLBs do documento d e

$W_{i,q}$ é o peso de termos e/ou RLBs da consulta q.

Estando definidos os termos e as RLBs e calculados os pesos, a classificação dos documentos depende do valor de relevância dos mesmos e da formulação Booleana da consulta.

3.6 Consulta Booleana

Uma consulta Booleana *qb*, conforme o modelo TR+, é formulada de acordo com a gramática apresentada a seguir, com formalismo BNF:

qb → *disjunçãoRLBs* OU (*conjunçãoTermos*)
disjunçãoRLBs → *r* OU *disjunçãoRLBs* | ε
conjunçãoTermos → (*disjunçãoTermos*) E *conjunçãoTermos* | ε
disjunçãoTermos → $\eta_1(p)$ OU $\eta_2(p)$
r → RLB
p → adjetivo | advérbio | particípio | substantivo | verbo
OU → operador Booleano de disjunção
E → operador Booleano de conjunção
ε → elemento vazio

Neste esquema, a conexão entre um descritor X e ε, através de um operador Booleano, tem o mesmo valor de X.

Se a consulta *q* em linguagem natural informada pelo usuário for, por exemplo, "pintura restaurada", então será formulada no formato Booleano, conforme o modelo TR+, a seguinte consulta *qb*:

$$r1 \text{ OU } r2 \text{ OU } ((\eta_1(p1) \text{ OU } \eta2(p1)) \text{ E } (\eta_1(p2) \text{ OU } \eta_2(p2)))$$

onde:

$r1$ = de(restauracao,pintura),
$r2 = r1'$ = ≠de(restauracao,pintura),
$\eta1(p1)$ = ε,
$\eta2(p1)$ = pintura,
$\eta1(p2)$ = restauracao,
$\eta2(p2)$ = restaurador,
$p1$ = pintura, e
$p2$ = restaurada.

A notação "≠de" significa qualquer identificador diferente de "de".

Os documentos recuperados são, então, classificados em dois grupos:

(i) grupo superior, de maior relevância: documentos que atendem às condições estabelecidas na consulta Booleana, ou seja, possuem pelo menos uma das RLBs da consulta ou, na falta de todas elas, possuem obrigatoriamente todos os termos conforme especificado; e

(ii) grupo inferior, de menor relevância: documentos que não atendem a todas as condições estabelecidas na consulta Booleana, mas possuem pelo menos um dos termos da consulta.

Em cada um desses dois grupos os documentos são classificados em ordem decrescente do valor de relevância.

No primeiro grupo, os documentos que possuem alguma RLB da consulta tendem a subir na classificação. Para cada RLB, no cálculo do valor de relevância, os termos presentes na RLB contribuem duplamente (como já foi mencionado): com seus próprios pesos e com o peso do relacionamento que formam. Se não há nenhuma RLB da consulta em um documento, ele tenderá a perder posições na classificação do primeiro grupo. Essa penalização se justifica porque os termos ocorrem no documento mas não se relacionam do modo como é especificado na consulta.

No segundo grupo, os documentos não possuem RLBs da consulta. Eles

possuem, no mínimo, 1 e, no máximo, n-1 dos n termos encontrados na consulta. É possível até mesmo que um documento deste grupo apresente valor de relevância maior que o de algum documento do primeiro grupo, em razão da elevada evidência de algum termo, mas ele será penalizado pela ausência de RLBs e, de qualquer forma, será classificado no grupo inferior.

É importante salientar que este esquema é específico para consultas curtas. Foi testado para consultas com até três termos. Para consultas longas, a exigência de que os documentos do primeiro grupo possuam todos os termos deve ser revista e, provavelmente, atenuada.

3.7 Resumo do Capítulo

Um texto não pode ser concebido apenas como uma seqüência de caracteres ou palavras. Além da freqüência de ocorrência, as palavras têm determinadas características morfológicas e relações de coesão úteis para o cálculo da representatividade dos descritores. No modelo proposto com dependência de termos para RI, denominado TR+, a evidência de um termo depende de sua participação em mecanismos de coesão frásica, e a evidência de um relacionamento depende das evidências de seus argumentos (termos).

No modelo TR+, os documentos e a consulta em linguagem natural recebem tratamento idêntico para a construção do espaço de descritores, no caso dos documentos, e para iniciar a formulação da consulta Booleana. Primeiramente, o texto é pré-processado com *toquenização* e etiquetagem morfológica; posteriormente é realizada a nominalização com a definição dos termos e, então, são identificadas as RLBs. Durante a busca, termos e RLBs da consulta são pesquisados no espaço de descritores e é calculado o valor de relevância para os documentos de acordo com o peso dos descritores e os operadores Booleanos predefinidos.

Operações de nominalização transformam uma palavra (adjetivo, advérbio, particípio ou verbo), existente no texto, em substantivos abstrato e/ou concreto, semanticamente correspondentes. RLBs identificam relacionamentos entre termos nominalizados, capturados de mecanismos de coesão frásica. Há RLBs de classificação, entre subclasses e classes de objetos; de restrição, entre modificado e modificador; e de associação, entre sujeito e objeto ou adjunto.

No modelo TR+, o grau de representatividade de um descritor é afetado (i) pelo processo de nominalização, (ii) pela capacidade das regras para identificação de RLBs de deduzir estruturas de dependência evidentes e (iii) pela formulação do cálculo do peso dos descritores, que leva em conta não só a freqüência de ocorrência mas, também, a ocorrência de mecanismos de coesão frásica.

A consulta Booleana, formulada através de disjunções de RLBs e conjunções de termos, complementa a especificação das dependências de termos e estabelece critérios para a classificação dos documentos por relevância.

No próximo Capítulo são descritas, visando a avaliação do modelo proposto: a coleção de documentos utilizada, as ferramentas desenvolvidas para implementar as diferentes estratégias comparadas, a metodologia de avaliação adotada e as estratégias avaliadas.

4 ASPECTOS METODOLÓGICOS E DE AVALIAÇÃO

4.1 Introdução

Sistemas de RI podem ser avaliados sob dois enfoques [BAE 99]: (i) avaliação de performance, quanto ao tempo gasto no processamento e quanto ao espaço de memória exigido, e (ii) avaliação de performance de recuperação, quanto à qualidade da saída de dados. A performance de recuperação é usualmente avaliada segundo diretrizes inicialmente estabelecidas no Projeto Cranfield [CLE 67]. Essas diretrizes foram ratificadas por importantes experiências através de avaliações dos sistemas SMART [SAL 68] e MEDLARS [LAN 69] e, posteriormente, consagradas pelas Text Retrieval Conferences (TRECs) [HAR 95, VOO 2003].

O paradigma Cranfield [BUC 2004] estabelece o uso de coleções de teste – ou coleções de referência –, com o objetivo de comparar a performance de recuperação de métodos alternativos para RI. Uma coleção de referência é constituída por (i) um conjunto de documentos, (ii) um conjunto de declarações de necessidade de informação (chamadas "tópicos" [VOO 2003]), e (iii) um conjunto de julgamentos de relevância, ou seja, listas de documentos que deveriam ser recuperados a cada tópico.

Seguindo esse paradigma, os experimentos de avaliação realizados neste trabalho necessitaram, então, de uma coleção de referência e da implementação do modelo proposto e de alternativas com as quais ele pudesse ser comparado. A coleção de referência utilizada é descrita na Seção 4.2, com informações sobre sua origem e suas características. Alternativas ao modelo proposto foram buscadas entre os trabalhos correlatos. Dois deles [SRI 2002, LEE 2005] foram reproduzidos conforme especificações disponíveis na bibliografia, outros seguiram características gerais dos modelos de dependência de termos estabelecidos na Seção 2.5. As estratégias alternativas analisadas são descritas na Seção 4.5.2.

Esses sistemas experimentais implementados utilizam diversos recursos que envolvem PLN. A aplicação de PLN nesses sistemas também foi avaliada. Essa avaliação sob a perspectiva do PLN (Capítulo 5) tem o objetivo de determinar a precisão dos recursos implementados em cada alternativa de RI. Ela pretende garantir a comparação justa dessas alternativas quanto à avaliação de performance de recuperação (Capítulo 6) e quanto à análise das relações custo/benefício envolvendo tempo de processamento e espaço de memória (Capítulo 7).

Visando a avaliação apresentada sob esses dois enfoques (PLN e resultados de RI) e a análise das relações custo/benefício, algumas tarefas prévias foram realizadas: (i) foi adaptada uma coleção de documentos, (ii) foram elaborados tópicos para formulação de consultas, (iii) foram desenvolvidas ferramentas para implementação das estratégias avaliadas e (iv) foram construídos os arquivos de índice correspondentes a essas estratégias.

No restante deste Capítulo, detalhes sobre a coleção de referência utilizada são informados, as ferramentas desenvolvidas são apresentadas e a metodologia de avaliação é descrita. Também são descritas as estratégias de indexação examinadas e é analisada a ocorrência dos três tipos de RLBs (classificação, restrição e associação) identificadas.

4.2 Coleção de referência Folha94

A coleção de referência Folha94 é constituída (i) por um conjunto de documentos, descritos na Seção 4.2.2; (ii) por 50 tópicos para formulação de consultas, apresentados no Anexo E; e (iii) pela listagem, apresentada no Anexo F, contendo os documentos relevantes para cada um dos tópicos de consultas utilizados.

A elaboração dos tópicos para formulação de consultas e a construção da relação dos documentos relevantes são assuntos da Seção 4.4.2. A origem do conjunto de documentos utilizados na coleção de referência Folha94 e as adaptações realizadas no corpus (coleção de documentos) de origem são explicadas a seguir.

4.2.1 Origem e preparação do conjunto de documentos utilizados

A coleção de documentos aqui utilizada foi adaptada a partir corpus FolhaNot[14]. O FolhaNot é constituído por 5.093 artigos de 229 edições do jornal Folha de São Paulo do ano de 1994. O texto, neste corpus, é verticalizado contendo, em cada linha, uma pontuação ou uma palavra com a correspondente etiqueta morfo-sintática. Um pequeno trecho inicial do FolhaNot é apresentado na Figura 4.1.

```
$START
<ed sec=agrofolha sem=94a data=05/04>
<artigo>
<t>
Barreiras     [barreira] N F P @SUBJ>
faz           [fazer] <fmc> V PR 3S IND VFIN @FMV
10            [10] <card> NUM M P @>N
anos          [ano] N M P @<ACC
com           [com] PRP @<ADVL
safra         [safra] N F S @P<
recorde       [recorde] ADJ F S @N<
</t>
<caixa>
Pólo          [pólo] N M S @SUBJ>
de            [de] <sam-> PRP @N<
...
```

Figura 4.1: Trecho inicial do corpus FolhaNot

Conforme pode ser observado neste trecho exemplificado, as palavras são armazenadas na forma original e na forma lematizada (entre colchetes). Outras marcações indicam início e final de artigo, título, parágrafo, frase, etc.

As seguintes alterações foram realizadas no FolhaNot para que fizesse parte da coleção de referência utilizada aqui, denominada Folha94:

- Redução do número de artigos. Foram descartados artigos incompletos (sem marca de final) e aqueles onde o corpo do artigo era formado por pequenas notas de jornal, com assuntos sem relação entre si, encabeçadas por mais de um título. Os 4.156 artigos restantes, numerados seqüencialmente, constituem os documentos da coleção de referência Folha94.

[14] O FolhaNot é um subconjunto etiquetado da parte jornalística do corpus NILC/São Carlos, compilado pelo NILC (http://www.nilc.icmsc.sc.usp.br), Núcleo Interinstitucional de Lingüística Computacional da Universidade de São Paulo (USP), em São Carlos. O FolhaNot foi etiquetado pela Linguateca (http://www.linguateca.pt) utilizando o *parser* PALAVRAS [BIC 2000].

- Eliminação das marcações (mantendo apenas as de artigo e de título) e transformação das etiquetas morfo-sintáticas em etiquetas morfológicas de forma a viabilizar as correções manuais explicadas a seguir.
- Correções da lematização. Por exemplo, o verbo "vira", lematizado em "ver", passou a ser lematizado, em alguns casos, em "virar".
- Correções de etiquetagem. Por exemplo, algumas ocorrências da palavra "meia" etiquetadas como adjetivos, passaram a ser etiquetadas como substantivos.
- Correções de locuções. Por exemplo, "para=casa", considerada locução adverbial, foi desmembrada em duas palavras: a preposição "para" e o substantivo "casa". Os substantivos compostos também foram desmembrados, embora alguns deles pudessem ser parcialmente aceitos. Por exemplo, "Jóquei=Clube=de=Sorocaba" foi desmembrado em quatro palavras.
- Informações acrescentadas. Foram incluídas informações decorrentes dos processos de *stemming* e nominalização aplicados a cada palavra.

4.2.2 Os documentos da coleção

O conjunto de documentos da coleção utilizada é constituído por 4.156 artigos de 229 edições do jornal Folha de São Paulo do ano de 1994. Cada artigo, daqui em diante, é identificado como um documento da coleção. Cada documento pode ser classificado em um ou mais tipos de assuntos. A Tabela 4.1 apresenta os tipos de assuntos encontrados na coleção.

Tabela 4.1: Tipos de assuntos presentes nos documentos da coleção

quantidade de documentos	tipos de assuntos
120	agropecuária e indústria alimentícia
1.142	cotidiano, polícia, ciência e educação
238	dinheiro e negócios
159	vida profissional
931	lazer e cultura
430	esporte
133	imóveis
503	informática e tecnologia digital
593	política
294	turismo e férias
126	veículos

Um pequeno trecho inicial da coleção de documentos utilizada é apresentado na Figura 4.2. Este trecho corresponde àquele do corpus FolhaNot apresentado na Figura 4.1. Na Figura 4.2 é possível observar a indicação de início do título (`<tit 1>`) e do texto (`<art 1>`) do primeiro documento. Apesar dessa marcação, neste trabalho, o texto de cada documento inclui, indistintamente, seu título.

Os documentos da coleção são armazenados em forma verticalizada (ver Figura 4.2), com cada item de texto[15] em uma linha. Cada linha é constituída pelo item de texto

[15] Será usada, neste trabalho, a expressão "item de texto" para identificar qualquer item encontrado no texto, como uma palavra, uma abreviatura, um símbolo (como o de tonelada) ou uma pontuação.

no formato original e, sendo palavra, na forma lematizada, na forma de *stem* e nas formas de substantivos abstratos e concretos (conforme o caso). Um sinal de igual (=) substitui um *stem* com o mesmo formato da palavra original e um zero (0) indica ausência de nominalização. Cada item de texto possui uma etiqueta morfológica, conforme a Tabela 4.2. Outro exemplo de trecho da coleção utilizada é apresentado no Anexo C.

```
<tit 1>
Barreiras barreira barreir 0 0 _SU
faz fazer faz 0 0 _VA
10 10 = 0 0 _NC
anos ano an 0 0 _SU
com com = 0 0 _PR
safra safra safr 0 0 _SU
recorde recorde record recorde 0 _AJ
. . = 0 0 _PN
<art 1>
Pólo polo pol 0 0 _SU
de de = 0 0 _PR
...
```

Figura 4.2: Trecho inicial da coleção de documentos utilizada

Tabela 4.2: Etiquetas, categorias morfológicas, palavras e pontuações

etiquetas	categorias morfológicas	palavras e pontuações			
		total	%	diferentes	%
_AD	artigos definidos	141.011	11,4	4	<0,1
_AI	artigos indefinidos	15.102	1,2	2	<0,1
_AJ	adjetivos	62.077	5,0	8.346	13,1
_AV	advérbios	47.848	3,9	1.234	1,9
_CC	conjunções coordenativas	30.155	2,4	12	<0,1
_CS	conjunções subordinativas	9.826	0,8	26	<0,1
_IN	interjeições	85	0,0	19	<0,1
_NC	numerais cardinais	32.580	2,6	4.575	7,2
_NO	numerais ordinais	822	0,1	102	0,2
_AP	particípios	23.440	1,9	4.940	7,8
_OS	pronomes possessivos	7.400	0,6	19	<0,1
_PD	pronomes demonstrativos	7.795	0,6	18	<0,1
_PI	pronomes indefinidos	12.712	1,0	90	0,1
_PL	pronomes relativos	13.481	1,1	13	<0,1
_PN	pontuações	71.473	5,8	10	<0,1
_PP	pronomes pessoais	15.778	1,3	29	<0,1
_PR	preposições	179.016	14,5	209	0,3
_SU	substantivos	334.782	27,1	29.760	46,8
_SU	substantivos *stopwords*	8.539	0,7	1.471	2,3
_VA	verbos auxiliares	37.474	3,0	213	0,3
_VB	verbos	86.197	7,0	12.476	19,6
_VG	vírgulas, parênteses e traços	97.698	7,9	4	<0,1
	total	1.235.291	100	63.572	100

Na Tabela 4.2, além das etiquetas utilizadas, é possível observar a quantidade de itens de cada categoria morfológica na coleção. Na coluna “diferentes”, os valores apresentados correspondem às ocorrências não duplicadas de cada categoria. Ocorrem apenas dois artigos indefinidos na coleção (“um” e “uma”) pois “uns” e “umas” são

sempre etiquetados indevidamente como pronomes indefinidos. São considerados "substantivos *stopwords*" palavras, como "sr." e "av.", etiquetadas como substantivos.

Na Tabela 4.3 são apresentadas, para cada categoria morfológica, as quantidades de termos derivados pelos processos de normalização lexical utilizados. Na coluna "termos diferentes derivados", foram desconsideradas as duplicatas.

Tabela 4.3: Termos derivados por categoria morfológica

categorias morfológicas	total de termos derivados				termos diferentes derivados			
	lemas	*stems*	abs	concr	lemas	*stems*	abs	concr
adjetivos	62.077 5,8%	61.339 7,7%	36.131 24,9%	12.660 11,1%	4.810 10,4%	4.121 13,2%	2.522 30,8%	714 12,0%
advérbios	47.848 4,5%	41.685 5,2%	5.924 4,1%	431 0,4%	1.217 2,6%	723 2,3%	468 5,7%	101 1,7%
particípios	23.440 2,2%	23.328 2,9%	22.766 15,7%	22.698 19,9%	2.317 5,0%	2.247 7,2%	2.191 26,8%	2.182 36,7%
substantivos	334.782 31,4%	330.520 41,2%	1.011 0,7%	2 <0,1%	29.393 63,7%	20.659 66,3%	58 <0,1%	2 <0,1%
verbos	86.197 8,1%	86.176 10,8%	79.300 54,6%	78.269 68,6%	3.051 6,6%	3.020 9,7%	2.944 36,0%	2.939 49,5%
outras	511.776 48,0%	258.308 32,3%	0 0,0%	0 0,0%	5.374 11,7%	371 1,2%	0 0,0%	0 0,0%
total	1.066.120 100,0%	801.356 100,0%	145.132 100,0%	114.060 100,0%	46.162 100,0%	31.141 100,0%	8.183 100,0%	5.938 100,0%

abs = substantivos abstratos
concr = substantivos concretos
As porcentagens são calculadas em relação aos valores da linha "total"

As derivações de substantivos nas colunas "abs" e "concr" se referem ao tratamento de sinonímia no processo de nominalização. Como, na grande maioria dos casos este tratamento foi aplicado aos substantivos abstratos, há apenas duas derivações de sinônimos para os substantivos concretos.

Entre os lemas estão incluídas, na Tabela 4.3, as 13.043 locuções (principalmente as adverbiais) que ocorrem nos documentos, repetidas como formas originais e normalizadas, conforme configuração do corpus FolhaNot. No caso de *stemming* e nominalização, processos executados posteriormente, o formato original não foi repetido e, assim, essas ocorrências não foram consideradas na quantidade de *stems* e formas nominalizadas.

Tabela 4.4: Substantivos originais e derivados na coleção

		derivados					
		presentes	ausentes do texto original			total	originais
substantivos	originais	no texto original	presentes no léxico	ausentes do léxico	total	de derivados	+ derivados
abstratos		5.716 69,9%	1.928 23,6%	539 6,6%	2.467 30,1%	8.183 100,0%	
concretos		2.975 50,1%	2.290 38,6%	673 11,3%	2.963 49,9%	5.938 100,0%	
total	29.760	8.691 61,5%	4.218 29,9%	1.212 8,6%	5.430 38,5%	14.121 100,0%	43.881

As porcentagens se referem ao total de derivados

Quanto à nominalização, ainda é preciso analisar as porcentagens de substantivos derivados. Na Tabela 4.4 são apresentados esses dados. As listagens de

substantivos abstratos e concretos obtidas por nominalização foram comparadas automaticamente com a listagem de substantivos presentes no dicionário Aurélio [FER 99]. Dessas comparações resultaram as informações apresentadas nas colunas "presentes no léxico" e "ausentes do léxico" da Tabela 4.4.

Entre os substantivos ausentes do léxico pode ser citado, como exemplo, o substantivo "acalentador", que não está presente em dois dicionários da língua Portuguesa: Aurélio [FER 99] e Houaiss [HOU 2002]. Este substantivo foi derivado segundo o padrão que determina que todo verbo terminado em "entar" tem substantivo concreto terminado em "entador", como "acrescentador", "afugentador" e "alimentador". Este padrão e todos os outros para verbos e adjetivos (assim como as exceções) foram pesquisados e confirmados no dicionário Aurélio. Todos os substantivos ausentes, derivados por nominalização, são gerados pelos autômatos de padrões (Seção 4.3.1).

É interessante observar que, ressalvados os erros do processo (Seção 5.3), a nominalização acrescenta informação não explícita no texto dos documentos. De todos os substantivos derivados, 38,5% não ocorrem no texto original dos documentos. Essa quantidade de informação é repassada à indexação pela nominalização.

Tabela 4.5: Espaço de memória utilizado pelos documentos da coleção

	espaço (Kb)	espaço (%)
itens de texto originais	7.153	24
lemas	6.855	23
stems	4.768	16
substantivos derivados	5.961	20
etiquetas morfológicas	5.067	17
total	29.804	100

Finalmente, resta examinar qual o custo de se armazenar outras informações junto ao texto dos documentos, além do próprio texto. Na Tabela 4.5 é apresentado o espaço de memória ocupado pelos itens de texto na forma original, e pelas informações oriundas de lematização, *stemming*, normalização e etiquetagem.

Optou-se por deixar armazenadas, junto aos documentos, as informações sobre a normalização lexical, porque os lemas já faziam parte do corpus FolhaNot e porque tais informações podem vir a ser úteis em trabalhos futuros.

4.3 Ferramentas desenvolvidas

Especificamente para a avaliação experimental do modelo TR+ foram desenvolvidas duas ferramentas: CHAMA, para o processo de nominalização, e RELLEX, para a identificação das RLBs. Outras ferramentas foram desenvolvidas visando a inclusão das demais estratégias na avaliação do modelo TR+. São as ferramentas para *stemming*, para captura de bigramas e para captura de sintagmas nominais.

Na Figura 4.3 é possível observar a dependência que existe entre as ferramentas utilizadas na construção dos espaços de descritores das estratégias de indexação avaliadas. Cabe lembrar que os procedimentos de *toquenização*, etiquetagem e lematização já haviam sido aplicados ao corpus FolhaNot, de onde se originou a coleção de documentos utilizada. Portanto, as ferramentas desenvolvidas utilizaram os dados produzidos por eles já armazenados na coleção de documentos junto ao texto. Não são apresentados na Figura 4.3, por motivos de simplificação, os procedimentos para

eliminação de *stopwords* e outros procedimentos para indexação, como inclusão dos descritores nos arquivos de índice.

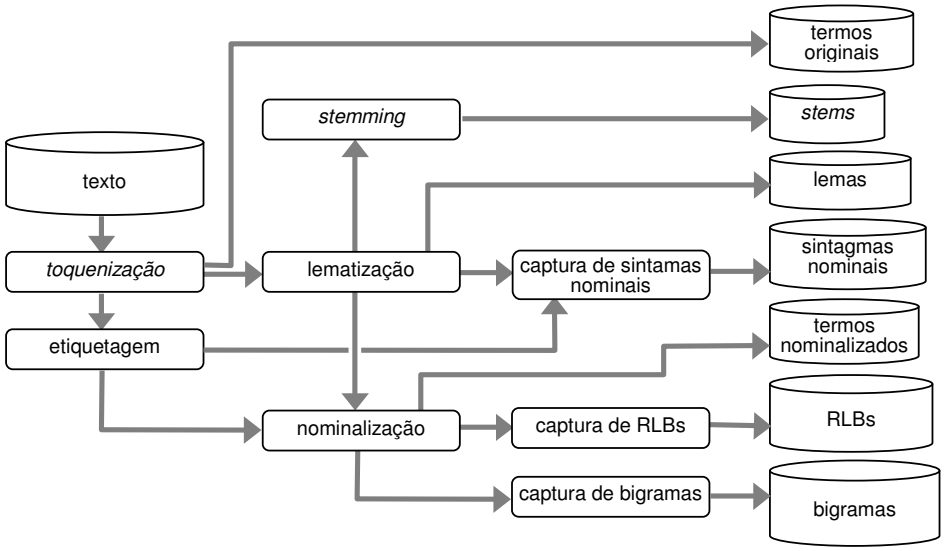

Figura 4.3: Ferramentas para PLN

Com exceção da ferramenta para *stemming*, que foi adaptada, as outras foram desenvolvidas inteiramente neste trabalho. Não houve adaptação nem foram utilizadas ferramentas prontas (i) porque não foram encontradas para a língua Portuguesa, no caso da nominalização, ou (ii) porque o desenvolvimento se mostrou mais adequado, no caso da identificação de RLBs. Há duas razões para esta decisão: (i) a interação necessária entre as ferramentas CHAMA e RELLEX (conforme será apresentado na Seção 4.3.3) e (ii) a preocupação com o tempo de execução. A ferramenta RELLEX executa algumas funcionalidades de um analisador sintático, somente aquelas necessárias para seu propósito. Um *parser* teria que ser adaptado para atender as regras para identificação das RLBs, sempre com a preocupação de otimizar o tempo de execução. Tal adaptação, provavelmente, demandaria tempo de programação maior que o gasto no desenvolvimento da ferramenta.

4.3.1 Ferramenta para nominalização

A ferramenta CHAMA foi desenvolvida para o Português, de acordo com o modelo TR+, para implementar de forma automática o processo de nominalização. Na Figura 4.4 é apresentada a estratégia adotada pela ferramenta.

A ferramenta CHAMA lê uma palavra lematizada[16] e retorna substantivos (abstrato e/ou concreto), quando há derivação válida. As palavras de entrada podem ser substantivos, adjetivos, advérbios e verbos (inclusive particípios). Os advérbios são transformados em adjetivos com a eliminação do sufixo "mente", quando o mesmo existe, e com a obtenção da forma masculina, quando o formato original for feminino. Por exemplo, "habilmente" é transformado em "hábil" e "remotamente" em "remoto" (neste caso a forma feminina – "remota" – foi alterada). Se não há sufixo "mente", a

[16] As palavras já se encontravam lematizadas no corpus FolhaNot. Isto simplificou o desenvolvimento da ferramenta.

forma original é mantida.

Os procedimentos de nominalização são aplicados através de três autômatos finitos: (i) AEx, para exceções, com 4.213 exceções, (ii) AAj, para padrões de adjetivos, com 660 padrões, e (iii) AVb, para padrões de verbos, com 350 padrões. Exceções e padrões são concebidos de acordo com as descrições contidas no Dicionário Eletrônico Aurélio [FER 99].

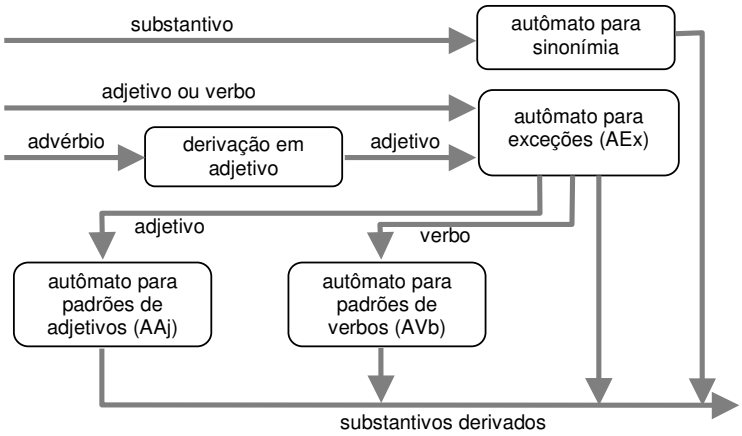

Figura 4.4: Estratégia da ferramenta CHAMA

Cada entrada de adjetivo ou verbo é analisada pelo autômato AEx para verificar se exige derivação irregular. Em caso afirmativo, a palavra é tratada como exceção e nominalizada desta forma. Se AEx rejeitar a palavra, é pesquisado um padrão de nominalização em AAj ou AVb.

Os três autômatos procuram pela maior cadeia de caracteres incluída na palavra. O autômato AEx lê a palavra do início para o final. Ao contrário, os autômatos AAj e AVb iniciam no caractere final e continuam em direção ao início da palavra. A leitura finaliza quando (i) a palavra inteira é lida ou (ii) quando não há transição válida para o próximo caractere a ser lido. Finalizada a leitura, a cadeia lida é (i) aceita, se o atual estado é um estado final do autômato, ou (ii) rejeitada, em caso contrário. Se a cadeia for rejeitada por AAj ou AVb, não há nominalização possível; se for aceita por um dos três autômatos, é realizada a nominalização.

Uma cadeia aceita pode ser a raiz (em AEx) ou o sufixo (em AAj ou AVb) da palavra de entrada, mas também outras partes podem ser incluídas para definir uma exceção ou um padrão. A cadeia aceita pode corresponder, até mesmo, à palavra inteira, se for necessário. É o caso do adjetivo "pirado", que é nominalizado de forma diferente ("piracao") de outros adjetivos com terminação "irado", como "peneirado" (cuja forma nominalizada é "peneiramento").

Cada estado final dos autômatos especifica duas operações de nominalização ($\eta 1$ e $\eta 2$) para derivar substantivos abstratos e concretos. Cada operação pode ter ações de remoção e/ou de inclusão de caracteres. Uma ação de remoção define quantos caracteres devem ser eliminados do final da cadeia lida. Uma ação de inclusão define novos caracteres que devem ser acrescentados ao final da cadeia lida. Na Tabela 4.6 são apresentados exemplos de operações.

Um estado final no autômato AEx pode apresentar, para adjetivos, alternativas diferentes das destinadas a verbos. Por exemplo, para derivar um substantivo concreto,

quando a cadeia "questi" é lida, o estado final indica a operação "+onador" para verbos e nenhuma operação para adjetivos (ver primeira e segunda linhas da Tabela 4.6).

Palavras com mesma terminação podem não ser tratadas da mesma maneira. Por exemplo, apesar de possuir o mesmo sufixo, os adjetivos "autoral", "normal" e "reverencial" sofrem diferentes operações para nominalização (ver Tabela 4.6). O adjetivo "autoral" é uma exceção reconhecida por AEx. Em AAj, a cadeia "al" define derivação para adjetivos como "normal", enquanto a cadeia "ncial" faz com que outro estado final seja alcançado no caso de adjetivos como "reverencial".

Tabela 4.6: Exemplos de operações para nominalização

palavra de entrada	autômato	cadeia	η_1	substantivo abstrato	η_2	substantivo concreto
questionar	AEx	questi	+onamento	questionamento	+onador	questionador
questionável	AEx	questi	+onamento	questionamento		ε
autoral	AEx	autoral		ε	–2	autor
normal	AAj	al	+idade	normalidade		ε
reverencial	AAj	ncial		ε	–1	reverencia
observar	AVb	ar	–1+cao	observação	–1+dor	observador
lembrar	AVb	embrar	–1+nca	lembranca	–1+dor	lembrador

ε = ausência de nominalização
\+ = indicativo de ação de inclusão
– = indicativo de ação de remoção

Verbos com a mesma terminação também podem ter nominalizações distintas. É o caso dos verbos terminados em "ar" na Tabela 4.6. O verbo "questionar" é aceito no autômato AEx, enquanto os verbos "observar" e "lembrar" se encaixam em padrões correspondentes às cadeias "ar" e "embrar", em AVb.

A ferramenta CHAMA inclui, também, um autômato para tratar da sinonímia de substantivos que constituem alternativas válidas de nominalização. Este autômato foi construído paralelamente à implementação dos três autômatos descritos anteriormente. Foi elaborada uma listagem de sinônimos da seguinte forma. Sempre que um substantivo devia ser selecionado entre outros para representar a nominalização de uma palavra (embora todos fossem válidos de acordo com o dicionário Aurélio), as alternativas foram registradas e um dos substantivos foi considerado preferencial. Por exemplo: "selecionamento" e "seleção" são alternativas válidas para o ato de "selecionar". No caso, foi considerado preferencial o substantivo "seleção", pois, no dicionário Aurélio, esta palavra é indicada como sinônimo de "selecionamento" e não vice-versa. Na ausência tal de indicação, a forma preferencial foi escolhida por ser a mais usual.

Tais alternativas deram origem às cadeias do autômato para sinônimos de substantivos. As operações de nominalização, neste autômato, derivam o substantivo preferencial em cada caso. Quando um substantivo original (da coleção de documentos) é aceito, o substantivo preferencial é derivado. Por exemplo, sempre que o substantivo "selecionamento" é encontrado, o sinônimo "seleção" é derivado. Este autômato possui 327 cadeias.

No Anexo A são apresentadas as cadeias e as operações que constam dos quatro autômatos utilizados.

4.3.2 Ferramenta para identificação de RLBs

A ferramenta RELLEX foi desenvolvida para o Português, de acordo com o modelo TR+, para identificação automática de RLBs. Na Figura 4.5 é apresentada a estratégia da ferramenta.

Para identificar RLBs em um texto, são reconhecidas frases, sendo estas estruturadas em sentenças, conforme a Figura 4.6. A sentença é o escopo da captura das RLBs. Uma sentença é constituída por um conjunto de frases, vinculadas entre si, finalizado por uma pontuação (exceto vírgula, parênteses ou travessão).

Uma frase pode ter todos ou alguns dos seguintes componentes:

- Lado esquerdo (LE): constituído por um sintagma nominal, que faz o papel de sujeito da frase. Pode incluir pronome relativo ou conjunção, ou pode, ainda, conter apenas um destes dois tipos de palavras.
- Lado direito (LD): constituído por um sintagma nominal ou preposicional (ou ambos), que faz o papel de objeto (direto ou indireto) ou de adjunto.
- Conjunto verbal (CV): constituído pelo predicado sem os constituintes do LD.

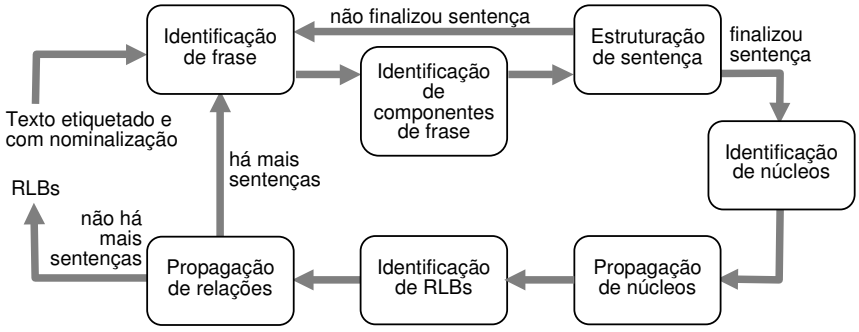

Figura 4.5: Estratégia da ferramenta RELLEX

Cada um destes componentes pode ter um núcleo. O núcleo do LE é o primeiro substantivo à esquerda. O núcleo do CV é último verbo à direita ou, na ausência dele, o último adjetivo ou particípio à direita. O núcleo do LD é o primeiro substantivo à esquerda não preposicionado. Os núcleos são utilizados pelas regras para identificação de RLBs apresentadas no Anexo B.

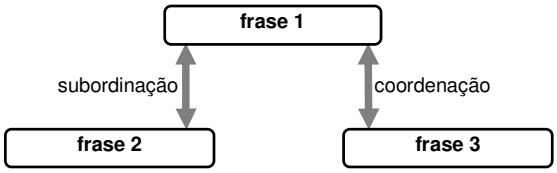

Figura 4.6: Exemplo de sentença estruturada em três frases

São considerados dois tipos de vínculos entre as frases de uma sentença: de coordenação ou de subordinação (ver Figura 4.6). Há procedimentos de propagação de núcleos e de RLBs que ocorrem através desses vínculos.

Por exemplo, a sentença "A fiel governanta, que trabalhou na casa de campo, e o mordomo fugiram" é decomposta nas seguintes frases:

Frase 1, com apenas o LE "A fiel governanta".

Frase 2, com o LE "que", o CV "trabalhou", e o LD "na casa de campo".

Frase 3, com o LE "e o mordomo" e o CV "fugiram".

Os componentes das três frases exemplificadas têm os núcleos sublinhados.

Uma frase iniciada por pronome relativo, conjunção subordinativa, preposição, parêntese ou traço é vinculada através de subordinação com a frase anterior (na mesma sentença), em caso contrário o vínculo é de coordenação. Esses vínculos tornam possível a propagação de núcleos e RLBs de uma frase para outra. O núcleo do LE da frase 1, ainda conforme o exemplo anterior, é propagado, em virtude do pronome relativo "que", para o LE da frase 2. Assim, "governanta" passa a ser, também, núcleo do LE da frase 2. A propagação de núcleos é uma forma de resolver alguns casos simples de co-referência.

As RLBs "de(fuga,*)", "por(fuga,*)" e "=(*,fugitivo)" são propagadas da frase 3 para a frase 1, devido à conjunção "e". Então, "governanta" e "mordomo" são argumentos válidos para compor essas RLBs, substituindo cada elemento representado pelo asterisco nas RLBs propagadas.

A ferramenta RELLEX utiliza 6 regras para identificar classificações, 15 regras para restrições e 3 regras para associações. As regras para identificação de RLBs são apresentadas no Anexo B.

Relacionamentos onde ocorrem advérbios de negação ou a conjunção "nem" não são tratados como RLBs. Quando isso acontece, é gerado um termo precedido por um hífen, denominado "termo negado". Por exemplo, "–fuga" é derivado de "não fugiu" enquanto "–mordomo" resulta de "nem o mordomo". Essa é uma tentativa de tratar, de forma simples e parcialmente, o complexo problema da negação.

4.3.3 Integração das ferramentas para nominalização e para identificação de RLBs

Conforme determina o modelo TR+, o processo de nominalização é realizado através de duas operações, apresentadas na Seção 3.3. Essas operações são exigidas porque, em alguns casos, há duas derivações necessárias, ou seja, podem existir dois substantivos distintos (um abstrato e outro concreto) relacionados semanticamente à palavra derivada. Existe, entretanto, outra razão: a ordem dos argumentos de uma RLB, em alguns casos, depende do tipo de substantivo derivado na nominalização. Por exemplo, nas expressões "praia tranquila" e "praia fluvial" são identificadas RLBs com os substantivos nominalizados em posições diferentes:

praia tranquila $\xrightarrow{rlb}$ de(tranquilidade,praia)

praia fluvial $\xrightarrow{rlb}$ de(praia,rio)

A diferença de posicionamento é possível porque são realizadas duas operações distintas em cada nominalização:

η_1(tranquila) = tranquilidade e η_2(tranquila) = ε

η_1(fluvial) = ε e η_2(fluvial) = rio

Em geral, a operação η_1 deve derivar substantivo abstrato e a operação η_2, substantivo concreto. Entretanto, há exceções. Uma operação η_2 pode derivar um substantivo abstrato. É o caso do adjetivo "publicitário", cuja nominalização estabelece:

η_1(publicitário) = ε e η_2(publicitário) = publicidade,

e não o contrário como deveria ser, conforme o modelo. Este artifício impede

campanha publicitária $\xrightarrow{rlb}$ de(publicidade,campanha),

e possibilita

campanha publicitária $\xrightarrow{rlb}$ de(campanha,publicidade)

Essas características são consideradas pelas ferramentas CHAMA, para nominalização, e RELLEX, para a identificação das RLBs.

4.3.4 Outras ferramentas

O modelo TR+ não necessita das ferramentas descritas nesta Seção e não as utiliza. Elas foram desenvolvidas apenas para a implementação de estratégias correlatas utilizadas na avaliação comparativa com o modelo proposto. As ferramentas desenvolvidas com este fim são: *stemmer*, ferramenta para captura de bigramas e ferramenta para captura de sintagmas nominais.

O algoritmo de *stemming* de Orengo e Huyck [ORE 2001] serviu de ponto de partida para o *stemmer* que foi desenvolvido aqui. Foram ampliadas as regras de remoção de sufixo, com a subdivisão em remoção irregular (com tratamento de exceções) e regular, que identifica padrões de sufixação. Para diminuir a complexidade da ferramenta, a entrada é constituída por palavras lematizadas.

A ferramenta para captura de bigramas emprega estratégia simples. São desconsideradas as *stopwords* do texto, processando apenas substantivos, adjetivos, advérbios, particípios e verbos, todos em forma nominalizada. Os bigramas são formados por pares de palavras adjacentes (descartadas as *stopwords*), não separadas por pontuação.

A ferramenta para captura de sintagmas nominais utiliza o módulo da ferramenta RELLEX que identifica os constituintes de cada sentença: o lado esquerdo (LE), o conjunto verbal (CV) e o lado direito (LD). Após o CV ser descartado, LE e LD sofrem o mesmo processamento: são selecionadas as formas lematizadas das palavras e eliminadas as palavras iniciais e finais que não são substantivos, adjetivos, advérbios ou particípios.

4.4 Metodologia de avaliação

A seguir são descritas as metodologias, utilizadas neste trabalho, para avaliação dos procedimentos para processamento da linguagem natural (PLN) e, também, para avaliação das estratégias de RI.

4.4.1 Avaliação dos procedimentos para PLN

Os procedimentos avaliados foram: etiquetagem de texto, normalizações lexicais (lematização, *stemming* e nominalização) e captura de relacionamentos (bigramas, sintagmas nominais e RLBs). Para cada um dos procedimentos avaliados foram coletadas 4 amostras. Cada uma delas foi analisada por avaliador humano sendo que cada um dos avaliadores humanos analisou, no máximo, uma amostra em cada procedimento avaliado.

Foi utilizada a estratégia de amostras aleatórias sistemáticas estratificadas [BIS 2004, CAL 2004]. Na amostragem estratificada os elementos são organizados por classes ou tipos, mantendo a proporção do todo na amostragem. É o caso, por exemplo, da amostragem das palavras por categoria morfológica.

Na amostragem aleatória sistemática os elementos da população estão (ou são) ordenados, a seleção do primeiro elemento se dá aleatoriamente, e a seleção dos próximos elementos ocorre a uma distância constante a partir do anterior de modo a se obter o número desejado de elementos da amostra. No caso da etiquetagem de texto e das normalizações lexicais, para realizar a amostragem, as palavras foram ordenadas alfabeticamente. No caso dos procedimentos de captura de relacionamentos, foi utilizada a ordem original de ocorrência das sentenças no texto dos documentos.

Nos casos da etiquetagem e das normalizações lexicais, foram descartadas as duplicatas para evitar a inclusão, na mesma amostra, de ocorrências idênticas e, em conseqüência, diminuir o tamanho do total a ser amostrado. Ao final, os erros encontrados foram ponderados levando em conta a freqüência de ocorrência das palavras com erro. A ponderação dos erros se dá conforme o seguinte exemplo. Há 62.077 ocorrências de adjetivos na coleção, sendo que 8.346 são diferentes. Se fossem analisados 100 adjetivos, descartando as duplicatas, a amostra teria 1,2% dos 8.346 adjetivos diferentes. Se fossem encontrados somente 2 erros de *stemming*, por exemplo:

stemming(carentes) = car e

stemming(precário) = prec,

seriam 2 erros em 100 adjetivos analisados, ou seja, 98% de precisão. Entretanto, na coleção há 8 ocorrências de "carentes" e 16 de "precário", o que resultaria em 24 erros no total. A ponderação, através de regra de três simples,

$$\frac{1{,}2}{100} = \frac{24}{x}$$

determinaria o valor de x = 2.000. Essa quantidade em 62.077 adjetivos corresponderia a 96,8% de precisão.

O descarte das duplicatas e a ponderação salientam os erros cometidos em palavras com elevada freqüência nos documentos, e menosprezam aqueles encontrados em palavras mais raras. Os resultados podem ser usados como uma estimativa dos erros existentes nos documentos da coleção utilizada, mas sobretudo visam avaliar o reflexo desses erros na RI. Um erro em algum termo da consulta, pode impedir ou forçar coincidência de descritores, afetando o cálculo da relevância. Quanto mais freqüente o termo com erro na coleção, maior o prejuízo na recuperação dos documentos.

No caso da etiquetagem, ao projetar as porcentagens para o total de erros estimados na coleção de documentos, foi tomado o cuidado de confirmar os erros com possível ambigüidade. Por exemplo, em algumas ocorrências "sentido" pode ser particípio, em outras, substantivo.

O tamanho de cada amostra foi dimensionado levando em conta os seguintes critérios: (i) o tamanho máximo limitado pela viabilidade do procedimento de análise humana, e (ii) o tamanho mínimo estabelecido com a intenção de atingir margem de erro admitida não superior a 2 pontos percentuais para mais ou para menos em relação à proporção observada num intervalo de confiança de 95%.

A margem de erro admitida é calculada através da aproximação quadrática para a obtenção do intervalo de confiança [CAL 2004]. Os limites inferior L_i e superior L_s são:

$$L_i = \frac{(2np + z^2 - 1) - z\sqrt{z^2 - (2 + 1/n) + 4p(nq + 1)}}{2(n + z^2)} \text{ e}$$

$$L_s = \frac{(2np + z^2 + 1) + z\sqrt{z^2 + (2 - 1/n) + 4p(nq - 1)}}{2(n + z^2)} ,$$

onde:

n é a quantidade de elementos da amostra,

$q = 1 - p$,

p é a proporção observada e

z é o valor crítico para a confiança desejada (z = 1,96 para intervalo de 95% de confiança).

4.4.2 Avaliação das estratégias de RI

Para a avaliação comparativa das estratégias de RI foi utilizada a coleção de referência Folha94, descrita na Seção 4.2. A metodologia de avaliação considerou a elaboração de tópicos para formulação de consultas, julgamentos de relevância e especificação de métricas para apresentar os resultados de recuperação.

Elaboração de tópicos de consultas

Foi adotada a metodologia utilizada nas TRECs (*Text Retrieval Conferences*) [VOO 2003], com tópicos para formulação de consultas. Um tópico é constituído por (i) um título, com as palavras que melhor descrevem o tópico; (ii) uma descrição, um parágrafo em linguagem natural descrevendo o tópico; e (iii) uma narrativa, uma relação concisa das características que identificam um documento relevante.

Os tópicos foram elaborados da seguinte maneira. Foram enviadas solicitações por correio-eletrônico a diversas pessoas para construção de consultas com os seguintes requisitos:

- objetivo: avaliação de sistema de recuperação de informação;
- tamanho da consulta: uma a três palavras; e
- contexto da pesquisa: artigos de jornal (Folha de São Paulo) do ano de 1994, com as seções ... (foram listados os tipos de assuntos apresentados na Tabela 4.1).

Foram descartadas as sugestões de consultas que não conseguiram recuperar documento algum da coleção utilizada. As primeiras 50 sugestões válidas geraram os títulos dos tópicos apresentados no Anexo E. Os itens descrição e narrativa foram elaborados, conforme a metodologia utilizada nas TRECs, com o objetivo de auxiliar o julgamento de relevância. As consultas formuladas, de acordo com cada estratégia avaliada, utilizaram apenas o título dos tópicos, ou seja, seguiram o modelo de consultas curtas (*short queries* [KWO 98, BRO 2003]).

Julgamentos de relevância

Foi adotado o método de *pooling* [BUC 2004], também utilizado nas TRECs, para o julgamento de relevância dos documentos recuperados pelas estratégias executadas para cada consulta. A seguir, são descritos os procedimentos executados.

Após a execução da busca e classificação para atender à consulta corrente, levando em conta os documentos recuperados por todas as estratégias, seguiram-se os passos para o julgamento de relevância:

(i) os 100 primeiros documentos recuperados em cada estratégia foram agrupados e ordenados, com eliminação das duplicatas;

(ii) a listagem resultante foi filtrada para descartar os documentos que, sem

dúvida, não continham informação relevante para a consulta;

(iii) a listagem final, quando formada por até 10 documentos, foi enviada ao formulador do tópico da consulta corrente; em caso contrário foi distribuída para até quatro pessoas (dependendo da extensão da listagem), sendo uma delas o autor do tópico;

(iv) cada julgador analisou a listagem que recebeu, marcando os documentos que considerou relevantes.

Para ser considerado relevante a uma dada consulta, um documento deveria conter alguma informação que merecesse (por critério do julgador) ser incluída em um relatório, caso este fosse realizado, sobre o tópico referente à consulta.

Com exceção da análise de relevância realizada por julgador humano, todos os outros procedimentos descritos anteriormente foram automatizados.

Uma das razões para a ordenação dos documentos recuperados (primeiro passo) foi a eliminação de qualquer possibilidade de vinculação dos documentos em julgamento a alguma estratégia utilizada. Esta medida visou tornar não tendenciosas a filtragem (segundo passo) e a marcação dos relevantes (quarto passo). A filtragem teve como objetivo a diminuição do tamanho da listagem final, tornando, assim, mais confiável a marcação dos relevantes. A filtragem foi realizada da seguinte maneira:

(i) o texto de cada documento da listagem foi apresentado com os termos da consulta destacados quando ocorressem;

(ii) o documento apresentado foi desconsiderado se, além de não ter termo algum destacado, também não houvesse dúvida quanto a sua não relevância para a consulta, ao serem analisados título, primeira e última frase; e

(iii) o documento apresentado foi mantido se contivesse algum termo destacado ou se, no caso anterior, existisse alguma dúvida.

Métricas

São utilizadas, na avaliação das estratégias de RI, as métricas precisão (*P*) e revocação (*R*)[17]:

$$P = \frac{rec\ \ rel}{rec} \text{ e}$$

$$R = \frac{rec\ \ rel}{rel},$$

onde:

rec rel é a quantidade de documentos recuperados relevantes;
rec é a quantidade de documentos recuperados; e
rel é a quantidade de documentos relevantes.

Também é utilizada, como média harmônica entre *P* e *R* , a medida *F* [BAE 99]:

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}}\ .$$

Outras medidas, como a "*mean uninterpolated average precision*" (MAP) [BUC 2004], foram testadas mas, como não alteraram significativamente os resultados das comparações realizadas e (algumas) são propostas muito recentes, não foram

[17] Conforme as métricas tradicionais, em Inglês, *precision* e *recall* [CLE 67, SAL 68, BUC 2004].

incluídas na avaliação.

4.5 Outras informações sobre a avaliação realizada

4.5.1 Notação para identificação das estratégias

As estratégias de indexação que usam unigramas são representadas através de duas letras maiúsculas:

- VR, com variantes das palavras originais,
- ST, com *stems*,
- LM, com termos lematizados, e
- NM, com termos nominalizados,

e as que consideram dependência de termos, através de quatro letras maiúsculas:

- SINN, com termos lematizados em sintagmas nominais,
- BIGR, com termos nominalizados em bigramas, e
- NMRL, com termos nominalizados e RLBs.

As identificações das estratégias de busca incluem as letras maiúsculas das correspondentes estratégias de indexação utilizadas, com a combinação de alguns dos seguintes afixos:

- prefixos:
  - "só": para estratégias que usam somente relacionamentos como descritores (não usam termos),
  - "bt": para a estratégia que implementa o trabalho correlato TCBT, descrito na Seção 2.6.9, de Shrikanth e Srihari [SRI 2002], com bitermos (bigramas sem restrição de ordem dos termos), e
  - "mm": para a estratégia[18] que implementa o trabalho correlato TCMM, descrito na Seção 2.6.10, de Changki Lee e Gary Lee [LEE 2005], com pares modificado-modificador em conexões gramaticais;
- sufixos:
  - "f": para as estratégias baseadas em freqüência de ocorrência,
  - "e": para as estratégias baseadas em evidência,
  - "s": para as estratégias que usam cálculo simples de peso de descritores, especificado quando é o caso,
  - "1": para as estratégias que usam consultas com apenas um termo,
  - "2": para as estratégias que usam consultas com dois ou mais termos, e
  - "B": para as estratégias com consulta Booleana (conforme Seção 3.6).

4.5.2 Descrição das estratégias de indexação

Foram examinadas sete estratégias de indexação, sendo quatro delas baseadas em unigramas: VR, LM, ST e NM, e três que consideram dependência de termos:

[18] No cálculo dos descritores desta estratégia foram adotados os parâmetros k_7=-5 e k_8=-0,1, pois propiciaram os melhores resultados.

BIGR, SINN e NMRL. Como exemplo, no Anexo C são apresentados os descritores correspondentes a um trecho de um documento da coleção utilizada, para cada uma das estratégias de indexação examinadas. Na Figura 4.7 é possível observar as transformações realizadas para a construção dos espaços de descritores de cada estratégia de indexação. Cada círculo com a notação da estratégia indica a origem de cada espaço de descritor. A eliminação de *stopwords* e de duplicatas de palavras (ou termos) é realizada em momentos diferentes e com objetivos diferentes em cada estratégia. Por exemplo, no caso dos *stems*, são eliminados *stems* duplicados; no caso dos sintagmas nominais e das RLBs, as preposições são consideradas na construção dos relacionamentos.

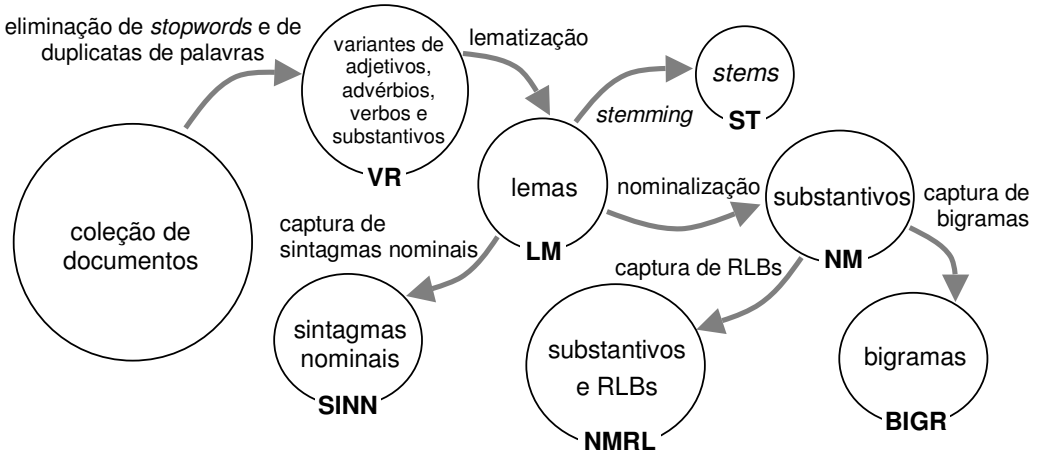

Figura 4.7: Transformações para construção dos espaços de descritores

Na Tabela 4.7 são caracterizadas resumidamente as estratégias de indexação examinadas, sendo informados quais os descritores utilizados e qual o modelo seguido (conforme modelos apresentados na Seção 2.5).

Tabela 4.7: Caracterização das estratégias de indexação

estratégias de indexação	descritores		modelo
	termos	relacionamentos	
VR	variantes originais	–	UG
LM	lemas	–	UG
ST	*stems*	–	UG
NM	nominalizados	–	UG
BIGR	nominalizados	bigramas	NG
SINN	lemas	sintagmas nominais	RT
NMRL	nominalizados	RLBs*	TT ou TR+

*TT usa RLBs apenas para o cálculo do peso dos termos

Todas as estratégias consideram substantivos, adjetivos, advérbios, particípios e verbos como palavras válidas (não *stopwords*) para a construção do espaço de descritores. As outras categorias morfológicas são tratadas como *stopwords*. Como já foi mencionado, nas estratégias SINN e NMRL, entretanto, para estruturar os relacionamentos, algumas *stopwords*, como as preposições, são também incluídas.

A estratégia VR utiliza, como descritores, termos constituídos pelas variantes

das palavras originais do texto, ou seja, no formato em que ocorrem. LM utiliza, como descritores, termos derivados de lematização das palavras do texto. Em ST, os descritores são termos na forma de *stems* das palavras do texto. Em NM, os descritores são termos derivados por nominalização.

As estratégias BIGR, SINN e NMRL incluem relacionamentos nos seus espaços de descritores. Em BIGR, os bigramas são constituídos por pares de termos nominalizados. SINN utiliza sintagmas nominais como descritores. NMRL adota o modelo TR+, ou seja, seus descritores são termos nominalizados e RLBs.

A derivação de *stems*, sintagmas nominais e substantivos nominalizados a partir de lemas foi uma decisão tomada em função de já existirem as formas lematizadas no corpus FolhaNot. Esta decisão tem uma vantagem e uma desvantagem. A vantagem foi a simplificação da implementação das ferramentas necessárias para as derivações a partir dos lemas. A desvantagem foi o risco de transferência dos erros de lematização para as derivações decorrentes.

Os sintagmas nominais foram derivados dos lemas e não dos termos nominalizados (como os bigramas), por duas razões: (i) para se aproximarem mais do formato como a consulta em linguagem natural é formulada, e (ii) para serem geradas duas alternativas de espaços de descritores (sintagmas nominais e bigramas) com maior número de diferenças entre si.

Todos os descritores foram gerados sem acentuação e em letra minúscula, inclusive no caso de VR.

4.5.3 Comportamento e constituição dos descritores

Os descritores constituídos apenas por termos não apresentam comportamento semelhante ao dos relacionamentos, quanto ao número de documentos onde ocorrem. Na Tabela 4.8 são apresentadas as proporções dos descritores quanto à freqüência por documento.

Os relacionamentos são descritores nitidamente mais específicos que os termos, isto é, são mais raros na coleção de documentos. A proporção de relacionamentos que ocorrem em apenas um documento na coleção, em BIGR, SINN e NMRL, é aproximadamente o dobro da proporção dos termos, em VR, LM, ST, NM e NMRL, conforme a Tabela 4.8. As RLBs do tipo associação são um exemplo extremo, pois não se repetem em mais de 10 documentos e 98,3% delas ocorrem em apenas um documento.

Tabela 4.8: Proporção de descritores por documento

estratégias		% de descritores que ocorrem em							quantidade de descritores
		1 doc.	2 docs.	3 docs.	4-5 docs.	6-10 docs.	11-100 docs.	>100 docs.	
VR		48,6	14,7	7,7	8,2	19,8	11,4	1,1	57.366
LM		47,6	13,8	7,5	7,8	21,2	12,9	2,2	37.267
ST		44,7	13,2	7,4	8.0	23,1	14,6	3,7	24.013
NM		47,2	13,5	7,5	7,6	21,5	13,4	2,8	36.479
BIGR		84,4	9,1	2,8	2,0	1,7	0,6	<0,1	296.243
SINN		88,1	5,9	1,9	1,6	2,5	1,2	0,1	112.239
NMRL	termos	47,5	13,8	7,4	7,7	21,0	12,9	2,6	38.616
	classificações	89,9	6,6	1,7	1,0	0,1	<0,1	<0,1	54.714
	associações	98,3	1,5	0,1	<0,1	<0,1	0	0	29.613
	restrições	89,0	7,3	1,9	1,2	0,1	<0,1	<0,1	160.766

Na Tabela 4.9 é apresentada a proporção dos descritores quanto à quantidade de componentes. Na coluna "média" aparece o número médio de componentes por descritor em cada estratégia. As estratégias VR, LM e NM, embora usem unigramas, não apresentam 100% de descritores com um termo em razão das locuções, como "chato=de=galocha" e "ainda=assim", que ocorrem neste formato nos documentos. O mesmo fato afeta o número de componentes dos termos na estratégia NMRL.

A estratégia BIGR, por usar bigramas, possui a maioria dos seus descritores (95,2%) armazenados com dois componentes. Os casos onde há apenas um componente são aqueles em que não há vizinho imediato (anterior e posterior) na mesma frase. As locuções dão origem aos casos com mais de dois componentes.

Tabela 4.9: Proporção de descritores quanto à quantidade de componentes

estratégias		% de descritores com							média
		1 comp.	2 comp.	3 comp.	4 comp.	5 comp.	6 a 10 comp.	> 10 comp.	
VR		99,2	0,5	0,2	0,1	0	0	0	1,012
LM		98,8	0,8	0,3	0,1	0	0	0	1,018
ST		100	<0,1	0	0	0	0	0	1,000
NM		98,7	0,8	0,3	0,1	0	0	0	1,018
BIGR		2,9	95,2	1,6	0,3	0,1	<0,1	0	1,995
SINN		13,3	20,8	14,0	18,7	10,8	19,4	3,0	4,065
NMRL	termos	98,8	0,8	0,3	0,1	0	0	0	1,018
	classificações	0	99,9	0,1	<0,1	0	0	0	2,001
	associações	0	0	98,7	0,5	0,7	0	0	3,020
	restrições	0	0	99,2	0,4	0,4	0	0	3,012

A estratégia SINN apresenta descritores com maior variação de número de componentes pois são armazenados na forma de sintagmas nominais. Há uma pequena preferência pela ocorrência desses descritores com dois a quatro componentes (53,5%).

No caso das RLBs, as relações dos tipos associação e restrição possuem, por definição, três componentes (dois argumentos e um identificador), enquanto as relações do tipo classificação possuem dois componentes (dois argumentos, sem considerar o identificador "=" como componente). Também aqui as locuções interferem na constituição desses descritores e, por esta razão, não há ocorrência de 100% para relações dos tipos associação e restrição com três componentes e para classificações com dois componentes.

4.5.4 Tipos de RLBs identificadas

As RLBs encontradas nos documentos apresentam a seguinte proporção: 65,6% são restrições, 22,3% são associações e 12,1% são classificações.

As restrições são constituídas pelos seguintes identificadores de relação:

- de: 57,9%,
- de/por: 14,3%,
- por: 9,8%,
- em: 8,0%,
- a: 3,0%,
- com: 2,5%,

- para: 1,9% e
- outros identificadores: 2,7%.

O identificador "de/por" foi utilizado para armazenar restrições que ocorrem no mesmo documento tanto com identificador "de" quanto com identificador "por". Esta decisão foi tomada para reduzir o espaço de memória utilizado.

Nos experimentos de avaliação realizados, em razão das consultas utilizadas, os tipos de RLBs não tiveram a mesma participação na obtenção dos resultados de recuperação das estratégias que utilizaram esses relacionamentos como descritores.

Das 50 consultas utilizadas, 33 têm mais de um termo e, é claro, somente estas permitiram a identificação de RLBs para serem usadas em fase de busca. A partir das consultas, foram identificadas três classificações, 33 restrições e nenhuma associação. As três classificações foram identificadas em três consultas (uma em cada), sem outras RLBs. Portanto, as classificações participaram efetivamente em três das 50 consultas realizadas. Nas outras 30 consultas foram encontradas somente restrições. Foi identificada uma restrição por consulta em 27 destas 30 consultas, e em cada uma das três restantes foram identificadas duas restrições. Logo, as restrições foram as RLBs mais utilizadas.

Nos documentos, conforme as porcentagens informadas anteriormente, a proporção "restrição:classificação:associação" é 5,4:1,8:1. Considerando o conjunto de consultas utilizado, entretanto, esta proporção é 11:1:0.

Uma causa simples impediu a ocorrência de RLBs do tipo associação: as consultas foram formadas por sintagmas nominais originados dos títulos dos tópicos, como consultas curtas. Como uma associação é uma relação formada entre um agente, do LE (lado esquerdo) de uma frase, e um tema, do LD (lado direito) da mesma frase, não basta um sintagma nominal para gerar uma associação. O mesmo fato diminuiu a proporção de RLBs do tipo classificação pois muitas delas são obtidas a partir da relação, em uma frase, entre o sujeito e a forma nominalizada correspondente ao agente do verbo.

A participação de RLBs do tipo associação só pode ser testada com consultas longas, com pelo menos uma frase completa. Esta avaliação não pertence ao escopo do presente trabalho, ficando como trabalho futuro a ser realizado (Seção 9.2).

Embora as RLBs dos tipos associação e classificação tenham apresentado menor participação que as restrições, conforme explicado, é importante lembrar que essas RLBs (assim como as restrições) foram utilizadas, de acordo com modelo proposto, para determinar a evidência dos descritores. Neste sentido, vale a proporção dos tipos de RLBs encontrada nos documentos para caracterizar suas participações no experimento realizado.

4.6 Resumo do Capítulo

A coleção de referência Folha94 é constituída por (i) um conjunto de documentos, formado por 4.156 artigos do jornal Folha de São Paulo do ano de 1994; (ii) um conjunto de 50 tópicos para formulação de consultas e (iii) a relação dos documentos relevantes a esses tópicos.

Os documentos são armazenados em forma verticalizada, tendo em cada linha uma palavra ou uma pontuação. As palavras aparecem no formato original, na forma lematizada, na forma de *stem* e nas formas de substantivos abstratos e concretos (quando é o caso). Cada pontuação ou palavra possui uma etiqueta morfológica.

Especificamente para o modelo TR+ foram desenvolvidas duas ferramentas: CHAMA, para o processo de nominalização, e RELLEX, para a identificação das RLBs. Outras ferramentas foram desenvolvidas visando a avaliação experimental do modelo proposto: são as ferramentas para *stemming*, para captura de bigramas e para captura de sintagmas nominais.

A avaliação dos procedimentos executados por essas ferramentas e dos resultados da etiquetagem de texto e da lematização, foi realizada através da análise de amostras aleatórias sistemáticas estratificadas. Cada amostra foi analisada por avaliador humano.

A metodologia de avaliação comparativa das estratégias de RI considerou (i) tópicos de consulta, para a formulação de consultas curtas, (ii) julgamentos de relevância, com método de *pooling*, e (iii) uso das métricas precisão, revocação e medida *F* para apresentar os resultados obtidos.

Foram examinadas sete estratégias de indexação, sendo quatro baseadas em unigramas: VR, com variantes das palavras originais, ST, com *stems*, LM, com termos lematizados, e NM, com termos nominalizados, e três estratégias que consideram dependência de termos: SINN, com termos lematizados em sintagmas nominais, BIGR, com termos nominalizados em bigramas, e NMRL (que implementa o modelo TR+), com termos nominalizados e RLBs.

Nos experimentos de avaliação realizados, em razão das consultas utilizadas, a proporção entre as RLBs dos tipos restrição, classificação e associação foi de 11:1:0, diferente do que ocorreu nos documentos, onde a proporção foi de 5,4:1,8:1. Isto aconteceu porque nenhuma RLB do tipo associação foi identificada a partir das consultas, por não formarem frases completas. As RLBs do tipo restrição, portanto, tiveram maior participação, como descritores, nos experimentos realizados, mas todas influenciaram o cálculo da evidência, conforme o modelo proposto.

5 AVALIAÇÃO SOB A PERSPECTIVA DO PLN

5.1 Introdução

O objeto de análise neste Capítulo é o conjunto de procedimentos para PLN executados sobre o texto dos documentos da coleção de referência Folha 94. Foram analisados os resultados da etiquetagem morfológica e da lematização, assim como os resultados obtidos pelas ferramentas desenvolvidas no presente trabalho para *stemming*, nominalização e captura de sintagmas nominais, bigramas e RLBs.

Esta análise foi realizada por dois motivos. Primeiro, porque esses dados e ferramentas foram utilizados na avaliação das estratégias de RI examinadas, podendo introduzir desvios importantes nos resultados daquelas estratégias. Segundo, porque os resultados desta análise podem caracterizar a coleção de documentos utilizada, quanto à ocorrência de erros contidos nas informações armazenadas, já que alguns dos procedimentos mencionados fizeram parte da construção da mesma.

Para facilitar a interpretação dos dados apresentados neste Capítulo, são informadas na Tabela 5.1 algumas medidas de avaliação encontradas na literatura para procedimentos de PLN.

Tabela 5.1: Precisão e revocação de procedimentos de PLN na literatura

procedimento	precisão (%)	revocação (%)
etiquetagem	87 a 98	
lematização	93 a 99	
stemming	81 a 96	
parsing	75 a 94	82 a 90
captura de termos compostos	73 a 77	95
captura de sintagmas nominais	86 a 96	92 a 97
captura de relações semânticas	55 a 88	50 a 89

A seguir são apresentados os resultados das avaliações realizadas, com informações sobre a amostragem, a margem de erro admitida, a precisão e algumas outras medidas consideradas de interesse.

5.2 Etiquetagem morfológica

Para a avaliação dos resultados da etiquetagem morfológica, foram analisadas amostras constituídas por 1.927 adjetivos, advérbios, particípios, verbos e substantivos (3,4% do total), 5.336 *stopwords* (100% do total) e 14 pontuações (100% do total), excluídas as duplicatas. Foram coletadas amostras para pontuações e para cada uma das categorias morfológicas (conforme Tabela 5.2). Os casos ambíguos (com possibilidade de classificação em mais de uma categoria morfológica) foram listados, após a primeira avaliação, juntamente com as palavras anteriores e posteriores, para novo julgamento.

A margem de erro admitida é de –0,6 a +0,4 pontos percentuais num intervalo de confiança de 95%. A Tabela 5.2 apresenta a precisão da etiquetagem do texto.

Tabela 5.2: Precisão da etiquetagem do texto

stopwords e pontuações	precisão (%)	não *stopwords*	precisão (%)
artigos	99,9	adjetivos	97,5
conjunções	99,9	advérbios	86,9
interjeições	98,8	particípios	96,6
numerais	99,9	substantivos	97,0
pontuação	100,0	verbos	90,5
preposições	99,9	média	93,5
pronomes	99,6		
verbos auxiliares	99,9		
média	99,9	média final	98,9

A precisão da etiquetagem das palavras que não são *stopwords* é afetada pelos valores mais baixos obtidos no caso dos verbos e, principalmente, dos advérbios. A etiquetagem das locuções adverbiais tem importante parcela de contribuição nesses resultados. Exemplos de erros de etiquetagem de verbos e advérbios são, respectivamente, "dança" (em "Dicionário de Balé e Dança", deveria ser substantivo) e "como" (no sentido de "porque" em "como a água não ...", deveria ser conjunção).

5.3 Normalização lexical

5.3.1 Lematização

Quanto à lematização, foram analisadas amostras constituídas por 1.319 palavras (2,3% dos adjetivos, advérbios, particípios, verbos e substantivos), excluídas as duplicatas. Nas amostras, as palavras originais (cada uma com o lema correspondente) foram agrupadas por categoria morfológica.

A margem de erro admitida é de –0,5 a +0,2 pontos percentuais num intervalo de confiança de 95%. A Tabela 5.3 apresenta a precisão da lematização.

Tabela 5.3: Precisão da lematização

categorias	precisão (%)
adjetivos	99,7
advérbios	99,9
particípios	100,0
substantivos	99,6
verbos	99,9
média	99,7

Alguns exemplos de erros de lematização são: "deputada" (deveria ser "deputado") para o adjetivo "deputada", "cedinho" (deveria ser "cedo") para o advérbio "cedinho", "gigantes" (deveria ser "gigante") para o substantivo "gigantes", e "sequestar" (deveria ser "sequestrar") para o verbo "sequestrar".

5.3.2 *Stemming*

Para avaliar o *stemmer* utilizado, foram analisadas amostras constituídas por 1.003 palavras (1,8% dos adjetivos, particípios, advérbios, verbos e substantivos), excluídas as duplicatas. Nas amostras, as palavras originais (cada uma com o *stem*

correspondente) foram agrupadas por categoria morfológica.

A margem de erro admitida é de –2,0 a +1,7 pontos percentuais num intervalo de confiança de 95%. A Tabela 5.4 apresenta a precisão do *stemming*.

Tabela 5.4: Precisão do *stemming*

categorias	precisão (%)
adjetivos	95,8
advérbios	94,4
particípios	99,8
substantivos	87,7
verbos	96,6
média	91,1

Alguns exemplos de erros de *stemming* são: "ir" (deveria ser "irad") para o adjetivo "irado", "estri" (deveria ser "estrit") para o advérbio "estritamente", "perpetua" (deveria ser "perpetu" ou "perpetuad") para o particípio "perpetuadas", "ag" (deveria ser "agent") para o substantivo "agentes", e "frauda" (deveria ser "fraud") para o verbo "fraudar". Uma importante causa de erros de *stemming*, no caso dos substantivos, são os nomes próprios.

5.3.3 Nominalização

Para a avaliação da ferramenta de nominalização, foram analisadas amostras constituídas por 1.130 adjetivos, particípios, advérbios e verbos (2,6% do total) e 45 substantivos processados por sinonímia (100% do total), excluídas as duplicatas. Nas amostras, as palavras originais (cada uma com os termos nominalizados correspondentes) foram agrupadas por categoria morfológica.

A margem de erro admitida é de –1,0 a +0,7 pontos percentuais num intervalo de confiança de 95%. A Tabela 5.5 apresenta a precisão da nominalização.

Tabela 5.5: Precisão da nominalização

categorias	precisão (%)			
	nominalizações corretas			palavras
	substantivos abstratos	substantivos concretos	total	nominalizadas corretamente
adjetivos	99,6	99,6	99,6	99,2
advérbios	99,6	99,1	99,4	98,7
particípios	96,5	96,6	96,5	95,2
verbos	99,3	99,0	99,1	98,4
média	99,0	98,8	98,9	98,1

A "nominalização" dos substantivos através de sinonímia não foi incluída na Tabela 5.5 porque apenas 45 substantivos (descartadas as duplicatas), como já exposto, foram processados. Como não foi constatado erro nesses 45 casos, a inclusão dos mesmos eleva artificialmente a precisão da nominalização, devido à ponderação do tamanho da amostragem (bastante pequeno) para o total de substantivos. A precisão para a nominalização é de 98,9%, levando em conta as nominalizações realizadas, e de 98,1%, levando em conta a quantidade de palavras com ambas as nominalizações (abstratas e concretas) corretas.

Tabela 5.6: Tipos de erros na nominalização

	substantivos abstratos (%)			substantivos concretos (%)		
	há válido		derivou,	há válido		derivou,
categorias	derivou diferente	não derivou	mas não há válido	derivou diferente	não derivou	mas não há válido
adjetivos	0,00	0,31	0,14	0,28	0,03	0,06
advérbios	0,00	0,38	0,02	0,00	0,87	0,00
particípios	0,64	2,91	0,00	1,27	1,45	0,64
verbos	0,20	0,54	0,00	0,49	0,49	0,05
média	0,12	0,78	0,07	0,37	0,75	0,11

Na Tabela 5.6 são apresentadas as porcentagens dos tipos de erros verificados na derivação de substantivos nominalizados abstratos e concretos. Os tipos de erros identificados são os seguintes:

- "derivou diferente": há substantivo válido correspondente à palavra original, mas a ferramenta derivou outro substantivo, diferente daquele que é válido, indevidamente;
- "não derivou": há substantivo válido correspondente à palavra original, mas a ferramenta não derivou substantivo algum; e
- "derivou, mas não há válido": não há substantivo válido correspondente à palavra original, mas a ferramenta derivou um substantivo indevidamente.

A ferramenta CHAMA, ao nominalizar as palavras dos documentos, derivou 14.121 substantivos e, destes, 8,6% não estão presentes no léxico (ver Tabela 4.4). Os erros estimados da ferramenta que interessam, neste caso, são dos tipos "derivou diferente" e "derivou, mas não há válido", conforme a Tabela 5.6. Como os valores da Tabela 5.6 abrangem todos os substantivos derivados, os erros devidos somente aos ausentes do léxico devem ser inferiores.

Dentre os substantivos incorretos derivados, é possível que alguns sejam encontrados no léxico (embora incorretos para a derivação corrente). Não foi possível obter esta informação: o léxico teria de ser consultado a cada erro, o que tornaria a avaliação muito lenta.

Alguns exemplos de erros de nominalização são: "resto" (deveria ser "restabelecimento") para o particípio "restabelecido", "dívida" (não deveria haver derivação) para o advérbio "deveras", e ausência de derivação para o adjetivo "brabo" (deveria ser "brabeza") e para o verbo "superfaturar" (deveria ser "superfaturamento").

Na Tabela 5.7 são apresentadas as freqüências de ocorrência médias das palavras com erro, considerando os procedimentos de normalização lexical. São apresentadas, também, as freqüências de ocorrência médias com que palavras se repetem nos documentos, em cada categoria morfológica não *stopword*. Por exemplo, na coleção de documentos, um mesmo adjetivo pode ser encontrado (com a mesma flexão), em média, em 7,44 ocorrências. Por outro lado, os adjetivos com erro na normalização são menos comuns. Por exemplo, entre os adjetivos com erro na lematização, um desses adjetivos pode ser encontrado (com a mesma flexão), em média, em 1,50 ocorrências, conforme a Tabela 5.7.

Tabela 5.7: Freqüência das palavras com erro de normalização lexical

categorias	freqüência média das palavras	freqüência média das palavras com erros		
		lematização	*stemming*	nominalização
adjetivos	7,44	1,50	4,06	2,07
advérbios	38,77	1,00	10,60	8,57
particípios	4,74	–	1,00	5,07
substantivos	10,99	1,59	24,78	1,80
verbos	6,91	1,00	37,50	0,00
média	9,67	1,52	18,62	3,59

É possível observar na Tabela 5.7 que o procedimento de *stemming* adotado comete erros em palavras mais comuns que a média no caso de substantivos (freqüência média é 10,99 e com erro é 24,78) e, principalmente, em verbos (freqüência média é 6,91 e com erro é 37,50). Nos procedimentos de lematização e nominalização, a freqüência média das palavras com erro é menor que a média geral, ou seja, os erros são cometidos em palavras mais raras na coleção de documentos, principalmente no caso da lematização.

Como já foi mencionado, os procedimentos de *stemming* e nominalização recebem palavras lematizadas para executar a derivação e, assim, são afetados por ela. Também são afetados pela etiquetagem do texto. Na Tabela 5.8 são apresentadas as origens dos erros da normalização lexical.

Tabela 5.8: Origem dos erros de normalização lexical

	origens dos erros (%)				
	etiquetagem	lematização	*stemming*	nominalização	total
lematização	0,1	0,2	0,0	0,0	0,3
stemming	0,1	0,4	8,4	0,0	8,9
nominalização	0,2	0,2	0,0	1,5	1,9

A etiquetagem tem alguma influência, embora pequena, nos erros de normalização lexical. A lematização afeta mais o processo de *stemming* do que o de nominalização. A maioria dos erros cometidos, entretanto, tem origem nos próprios procedimentos. A pequena influência da etiquetagem e da lematização se justifica porque esses procedimentos tem elevada precisão, conforme já foi exposto.

5.4 Captura de relacionamentos

5.4.1 Bigramas

Foram analisadas amostras constituídas por 0,05% (1.480 relacionamentos) dos bigramas capturados nos documentos. Os trechos de texto de onde foram capturados os bigramas foram incluídos, na amostragem, com quatro palavras: duas do relacionamento e mais uma anterior e uma posterior.

A margem de erro admitida é de –0,4 a +0,1 pontos percentuais num intervalo de confiança de 95%.

A precisão da captura dos bigramas é de 99,9%. Os erros encontrados têm origens com a mesma proporção de ocorrência: (i) identificação de bigramas com

termos que deveriam estar separados por pontuação e (ii) inclusão indevida de *stopword* em bigramas, em razão de etiquetagem morfológica incorreta.

5.4.2 Sintagmas nominais

Para a análise da captura de sintagmas nominais (SNs), foi utilizado um conjunto de amostras constituídas por 0,4% das sentenças identificadas nos documentos, o que corresponde a 2.858 sentenças. De cada uma dessas sentenças foi selecionada, aleatoriamente, uma frase. Foram analisados os SNs capturados das frases selecionadas.

A margem de erro admitida é de –0,8 a +0,6 pontos percentuais num intervalo de confiança de 95%. Na Tabela 5.9 são apresentadas a precisão e a revocação na captura dos SNs levando em conta as seguintes fórmulas:

precisão = SNs corretos capturados / SNs capturados e

revocação = SNs corretos capturados / SNs corretos.

Tabela 5.9: Precisão e revocação na captura de SNs

precisão (%)	revocação (%)
96,4	99,7

Na Tabela 5.10 são apresentados alguns dados sobre a captura dos SNs. São chamados, aqui, SNs compostos aqueles que poderiam ser desmembrados em dois ou mais. Embora agrupados, na Tabela 5.10, junto aos SNs incorretos, os SNs compostos não devem ser computados como erros por duas razões: (i) tais ocorrências podem ser consideradas SNs, e (ii) não influenciam os resultados de recuperação na avaliação realizada, pois um SN de uma consulta pode ser encontrado como componente de um SN composto.

Tabela 5.10: Dados sobre a captura dos SNs

SNs	%
não identificados	0,3
incorretos	3,6
compostos	7,9

Na Tabela 5.10, a porcentagem dos SNs não identificados é calculada em relação ao número total de SNs corretos (capturados ou não). Quanto aos SNs incorretos e compostos, leva-se em conta o total de SNs capturados.

Um exemplo de SN composto é "quinto depoimento na corregedoria do departamento de inquérito policial do tribunal de justica de São Paulo". Alguns exemplos de erros na captura de SNs são: "ajuda eleitoral" é um SN não identificado ("ajuda" recebeu etiqueta de verbo) e "ar fanzine" é um SN incorreto capturado de "tirou do ar Fanzine".

5.4.3 RLBs

Para a análise da identificação de RLBs, foi utilizado o mesmo conjunto de amostras dos SNs. Foram analisadas 2.858 sentenças (0,4% do total). De cada uma dessas sentenças foi selecionada aleatoriamente uma frase. Foram analisadas as RLBs identificadas nas frases selecionadas.

A margem de erro admitida é de –1,2 a +1,1 pontos percentuais num intervalo de confiança de 95%. Na Tabela 5.11 são apresentados os valores de precisão e de revocação para a identificação das RLBs. Esses valores são obtidos a partir das seguintes fórmulas:

precisão = RLBs corretas identificadas / RLBs identificadas e

revocação = RLBs corretas identificadas / RLBs corretas.

Tabela 5.11: Precisão e revocação na identificação de RLBs

	precisão (%)	revocação (%)
parcial	91,0	93,4
geral	88,8	91,5

Na Tabela 5.11, o cálculo da precisão geral considera como corretas as RLB com todos os componentes corretos (identificador da relação e argumentos). No cálculo da precisão parcial o identificador da relação é desconsiderado. No cálculo da revocação parcial são consideradas corretas somente as RLBs evidentes. No cálculo da revocação geral são consideradas corretas também as não evidentes.

As RLBs podem ser evidentes ou não evidentes, conforme já foi exposto. As RLBs corretas não evidentes são as que necessitam informação semântica para serem identificadas. Por exemplo, em "representante do partido em Alagoas", a RLB "de(representante,partido)" é evidente. Entretanto, "em(representante,alagoas)" e "em(partido,alagoas)" são RLBs não evidentes, embora uma delas deva ser correta dependendo do contexto.

A Tabela 5.12 apresenta algumas observações sobre a identificação de RLBs. As porcentagens das RLBs não identificadas são calculadas em relação ao total de RLBs corretas (evidentes ou não). As outras porcentagens da mesma Tabela levam em conta o total de RLBs identificadas.

Tabela 5.12: Observações sobre a identificação das RLBs

RLBs		causa (%)			total (%)
		gramática	nominali-zação	identi-ficação	
não identificadas	não evidentes	0,0	0,0	6,6	6,6
	evidentes	0,2	0,0	1,7	1,9
com argumentos invertidos		0,0	0,3	0,3	0,6
com *id* incorreto e argumentos corretos		0,0	0,0	2,2	2,2
com *id* e argumentos incorretos		1,2	0,0	7,3	8,4

id = identificador de relação

Conforme a Tabela 5.12, erros gramaticais (como pontuações indevidas) encontrados no texto dos documentos tem pequena participação no caso de RLBs evidentes não identificadas e RLBs com *id* e argumentos incorretos. Inversões não realizadas entre nominalizações concretas e abstratas (ver Seção 4.3.3) têm metade da responsabilidade sobre os erros de RLBs com argumentos invertidos. No total, entretanto, boa parte dos erros ocorreu porque o processo de identificação de RLBs não previu os casos de exceção que originaram tais erros.

Um exemplo de RLB não evidente não identificada é "de(modo,fazendeiro)" (em

"que mostra o modo de vida dos antigos fazendeiros"). Alguns exemplos de erros na identificação de RLBs são: "=(filho,cego)" é uma RLB evidente não identificada por falha na propagação de núcleo do lado direito de frase (em "tem filho que é cego"), a RLB "de(coletividade,decisao)" tem argumentos invertidos (identificada em "decisões coletivas"), a RLB "de(motivo,aposta)" tem *id* incorreto (deveria ser "para") e argumentos corretos (em "... têm motivos para apostar que ..."), e a RLB "de(descontentamento,jornal)" tem *id* e argumentos incorretos (em "distribuidores de jornal descontentes").

Os erros de identificação de RLBs, quanto ao tipo (classificação, restrição ou associação), têm proporção semelhante à própria proporção de ocorrência dos tipos de RLBs.

5.5 Resumo do Capítulo

Os procedimentos para PLN executados sobre o texto dos documentos da coleção de referência Folha 94 foram avaliados porque (i) são possíveis causas de introdução de erros nas estratégias de RI examinadas nesta tese e (ii) dão uma boa idéia sobre a precisão do conteúdo dos documentos.

A etiquetagem morfológica para adjetivos, advérbios, particípios, verbos e substantivos tem precisão de 93,5%, e para *stopwords* e pontuações, 99,9%. No geral, a etiquetagem do texto apresenta uma precisão de 98,9%.

A lematização de adjetivos, advérbios, particípios, verbos e substantivos tem precisão de 99,7%.

A precisão do *stemming* de adjetivos, advérbios, particípios, verbos e substantivos é de 91,1%.

Constata-se que 98,1% dos adjetivos, advérbios, particípios e verbos apresentam ambas as nominalizações (abstrata e concreta) corretas. A precisão da nominalização, levando em conta a derivação de substantivos abstratos, é de 99,0% e, considerando a derivação de substantivos concretos, é de 98,8%. Os tipos de erros são, em ordem decrescente de ocorrência: (i) nenhum substantivo derivado quando há nominalização válida, (ii) substantivo derivado diferente do correto e (iii) substantivo derivado quando não há nominalização válida.

Em geral, os erros de *stemming* acontecem com palavras mais comuns (freqüência média de ocorrência = 18,62) nos documentos. Por outro lado, há uma tendência dos erros de lematização e de nominalização serem encontrados em palavras mais raras, respectivamente com freqüência média de ocorrência de 1,52 e 3,59. As palavras que não são *stopwords* têm freqüência média de ocorrência de 9,67 no texto da coleção de documentos utilizada.

A precisão na captura dos bigramas é de 99,9%, na captura dos sintagmas nominais é de 96,4%, e na identificação das RLBs é de 88,8%. Foram capturados 99,7% dos sintagmas nominais e 91,5% das RLBs existentes nos documentos.

6 AVALIAÇÃO SOB A PERSPECTIVA DA RI

6.1 Introdução

Neste Capítulo, são apresentadas avaliações comparativas entre as estratégias de busca examinadas. Essas avaliações permitem observar a influência, nos resultados da recuperação, de:

- diferentes processos de normalização lexical,
- diferentes modos de especificação de dependência de termos,
- representatividade dos descritores calculada com base em freqüência de ocorrência e em evidência,
- tamanho da consulta e
- presença/ausência de operadores Booleanos na consulta.

As estratégias de busca utilizadas para essas avaliações estão caracterizadas na Tabela 6.1. Cada uma delas está associada à correspondente estratégia de indexação. As demais colunas da Tabela são explicadas a seguir.

Na coluna "evidência" da Tabela 6.1, a presença da palavra "sim" significa que o cálculo de peso utilizado é baseado em evidência, e sua ausência significa que o cálculo é baseado em freqüência de ocorrência.

Na coluna "descritor" aparece a indicação de uso apenas de termos (t), nas estratégias com unigramas, ou de uso de termos e relacionamentos (t + r) ou apenas relacionamentos (r), naquelas estratégias que consideram dependência de termos.

Na coluna "peso do descritor", da Tabela 6.1, são indicadas as alternativas probabilística (Eq. 2), de modelagem de linguagem (Eq. 5) e probabilística baseada em evidência, adotada pelo modelo proposto (Eqs. 7, 8 e 9).

São consideradas, ainda, abordagens onde o peso $W_{i,d}$ do descritor i em um documento d é simplesmente dado por:

$$W_{i,d} = w_{i,d} \tag{12}$$

sendo:

$$w_{i,d} = f_{i,d} \tag{13}$$

ou $w_{i,d}$ é igual à evidência do descritor, obtida através da Equação 8. Deste modo, é evitada a interferência de qualquer outro fator no cálculo do peso dos descritores, além da própria freqüência de ocorrência ou da evidência.

Na coluna "operador Booleano", da Tabela 6.1, a presença da palavra "sim" significa uso de consulta com operador Booleano, e sua ausência significa que este tipo de operador não é utilizado.

Na coluna "consulta" é indicado se a consulta é formulada com apenas um termo (1 t), com dois ou mais termos (2 ou + t), ou se não há restrição do número de termos da consulta (caso representado pela ausência de indicação).

Tabela 6.1: Caracterização das estratégias de busca

estratégias de busca	estratégias de indexação	evidência	descritor	peso do descritor	operador Booleano	consulta
VRf	VR		t	Eq. 2		
STf	ST		t	Eq. 2		
LMf	LM		t	Eq. 2		
NMf	NM		t	Eq. 2		
NMf1	NM		t	Eq. 2		1 t
NMsfB	NM		t	Eq. 12 e 13	sim	
NMfB	NM		t	Eq. 2	sim	
NMfB1	NM		t	Eq. 2	sim	1 t
NMfB2	NM		t	Eq. 2	sim	2 ou + t
BIGRf	BIGR		t + r	Eq. 7 e 13		
BIGRfB	BIGR		t + r	Eq. 7 e 13	sim	
sóBIGRf	BIGR		r	Eqs. 7 e 13		
btBIGRf	BIGR		r	Eq. 5		
SINNf	SINN		t + r	Eq. 7 e 13		
SINNfB	SINN		t + r	Eq. 7 e 13	sim	
sóSINNf	SINN		r	Eq. 2		
mmNMRLf	NMRL		t	Eqs. 2 e 6		
NMe	NMRL	sim	t	Eqs. 7 e 8		
NMe1	NMRL	sim	t	Eqs. 7 e 8		1 t
NMeB	NMRL	sim	t	Eqs. 7 e 8	sim	
NMseL	NMRL	sim	t	Eqs. 12 e 8	sim	
sóRLBe	NMRL	sim	r	Eqs. 7 e 9		
NMRLeB	NMRL	sim	t + r	Eqs. 7, 8 e 9	sim	
NMRLe	NMRL	sim	t + r	Eqs. 7, 8 e 9		
NMRLeB1	NMRL	sim	t + r	Eqs. 7, 8 e 9	sim	1 t
NMRLeB2	NMRL	sim	t + r	Eqs. 7, 8 e 9	sim	2 ou + t

As estratégias de busca da Tabela 6.1 são agrupadas nas avaliações apresentadas a seguir. Nessas avaliações são apresentados gráficos com curvas revocação-precisão e tabelas que resumem os resultados da recuperação.

Nos gráficos, o eixo das abscissas corresponde à revocação (*R*) e o eixo das ordenadas à precisão (*P*). Cada ponto de uma curva nesses gráficos representa a precisão de uma estratégia correspondente a um dos 11 pontos de revocação (entre 0,0 e 1,0).

Nas tabelas são apresentados: valores de precisão para o documento do topo e para os primeiros 10 documentos; valores de revocação para os 10 e para os 100 primeiros documentos; e medida *F* para os 10 primeiros documentos classificados em cada estratégia. Através dos gráficos se tem uma visão geral dos resultados de recuperação, enquanto os dados das tabelas se concentram na porção superior da classificação por relevância, principalmente quanto aos 10 primeiros documentos recuperados.

6.2 Normalização lexical

O objetivo desta avaliação é comparar estratégias com diferentes processos para normalização lexical. São avaliadas as estratégias NMf, LMf, STf e VRf:

- NMf usa nominalização como processo de normalização lexical,
- LMf usa lematização,
- STf usa *stemming*, e

- VRf não usa processo de normalização lexical (os termos são indexados no formato como ocorrem no texto).

Todas as estratégias mencionadas adotam a Equação 2 para o cálculo do peso dos descritores, e, portanto, são baseadas em freqüência de ocorrência. Essas estratégias utilizam unigramas, ou seja, não há relacionamentos entre os descritores.

Na Figura 6.1 são apresentadas as curvas revocação-precisão para as estratégias NMf, LMf, STf e VRf. É possível observar a superioridade da nominalização (utilizada na estratégia NMf) sobre os outros processos de normalização lexical.

As curvas das estratégias LMf e STf se intercalam com valores superiores de precisão de LMf, quando a revocação está abaixo de 0,5, e com valores superiores de precisão de STf, quando a revocação está acima desse valor.

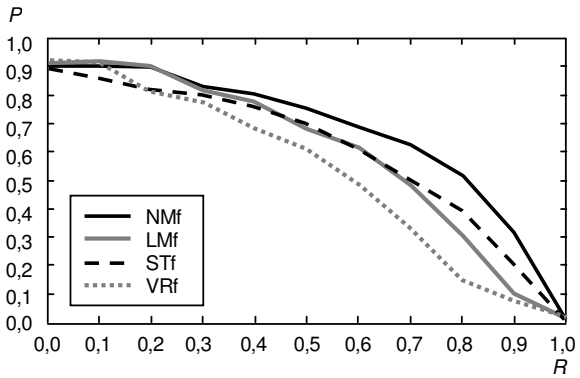

Figura 6.1: Curvas revocação-precisão para alternativas de normalização lexical

VRf apresenta a pior performance, entretanto tem bons resultados para o documento do topo (ver Tabela 6.2).

Na Tabela 6.2 são apresentados alguns valores de precisão, revocação e medida *F* confirmando a superioridade da estratégia NMf sobre as demais. As estratégias, na Tabela 6.2, são classificadas por ordem decrescente de performance.

Tabela 6.2: Alguns resultados para alternativas de normalização lexical

	precisão		revocação		*F*
estratégia	1	1–10	1–10	1–100	1–10
NMf	0,900	0,692	0,636	0,953	0,663
LMf	0,920	0,640	0,583	0,881	0,610
STf	0,900	0,626	0,587	0,934	0,606
VRf	0,920	0,600	0,529	0,808	0,562

1 = documento do topo, 1–N = primeiros N documentos

Nas condições desta avaliação, os resultados obtidos demonstram que a nominalização é uma alternativa interessante à lematização e ao *stemming*, pois a estratégia que a utiliza apresentou, no geral, valores de precisão, revocação e medida *F* superiores aos das estratégias com procedimentos tradicionais de normalização lexical.

6.3 Evidência e freqüência de ocorrência

Nesta avaliação, o objetivo é comparar estratégias com cálculos baseados em

freqüência de ocorrência e em evidência para o peso dos descritores. São avaliadas as estratégias NMe1, NMf1, NMseB e NMsfB:

- NMe1 usa as Equações 7 e 8, para cálculo (baseado em evidência) do peso dos termos, e consultas com apenas um termo;
- NMf1 usa a Equação 2 (baseada em freqüência) e consultas também com apenas um termo;
- NMseB usa a Equações 12 e 8 (baseadas em evidência) e adota consulta Booleana com um a três termos; e
- NMsfB usa as Equações 12 e 13 (baseadas em freqüência) e adota consulta Booleana, também, com um a três termos.

Nenhuma das estratégias desta avaliação usa relacionamentos como descritores, mas NMe1 e NMseB, naturalmente, consideram as RLBs para o cálculo da evidência.

O uso de operadores Booleanos por NMseB e NMsfB garante que os documentos com todos os termos da consulta são classificados por essas estratégias em posições superiores. Assim, a comparação entre freqüência de ocorrência e evidência é realizada com a presença de todos os termos nesses documentos, quando são recuperados documentos com essas características. Nessas duas estratégias nenhum outro fator, a não ser a freqüência e a evidência, é utilizado para o cálculo do peso dos descritores.

Foram executadas 17 consultas (do conjunto de 50) para NMe1 e NMf1. As consultas com mais de um termo foram eliminadas, neste caso, para garantir que NMe1 não tivesse nenhuma vantagem sobre NMf1 pela presença de mais termos. Essa preocupação se justifica porque o conceito de evidência leva em conta a coesão entre os termos. NMseB e NMsfB, ao contrário, usam todas as consultas com um a três termos.

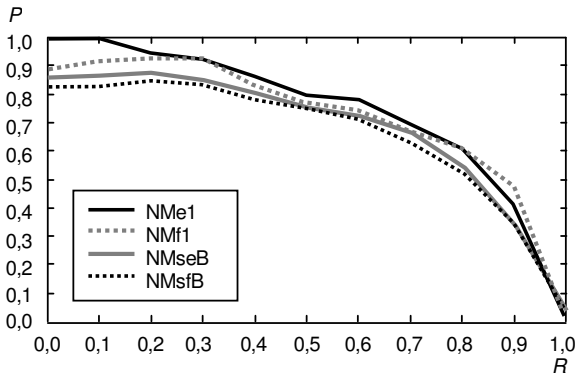

Figura 6.2: Curvas revocação-precisão para estratégias baseadas em freqüência de ocorrência ou em evidência

Salienta-se que há, aqui, duas comparações distintas: (i) entre NMe1 e NMf1 e (ii) entre NMseB e NMsfB. Qualquer outra comparação deve levar em conta que foram utilizados dois conjuntos diferentes de consultas.

Na Figura 6.2 é possível observar a superioridade do cálculo do peso dos descritores baseado em evidência sobre o cálculo baseado em freqüência de ocorrência: NMe1 apresenta resultados superiores a NMf1, assim como NMseB, com resultados superiores a NMsfB. Há superioridade da evidência tanto em consultas com apenas um termo quanto em consultas com mais termos. Na Tabela 6.3 também é possível observar

a superioridade de NMe1 sobre NMf1 e de NMseB sobre NMsfB.

Tabela 6.3: Alguns resultados para estratégias baseadas em evidência ou freqüência de ocorrência

	precisão		revocação		*F*
estratégia	1	1–10	1–10	1–100	1–10
NMe1	1,000	0,782	0,602	0,960	0,680
NMf1	0,882	0,747	0,572	0,960	0,648
NMseB	0,860	0,694	0,654	0,951	0,673
NMsfB	0,820	0,680	0,644	0,958	0,662

1 = documento do topo, 1–N = primeiros N documentos

Nas condições desta avaliação, os resultados obtidos demonstram que estratégias com cálculo do peso dos descritores baseado em evidência apresentam valores de precisão, revocação e medida *F* superiores àqueles obtidos por estratégias com cálculo do peso dos descritores baseado em freqüência de ocorrência.

6.4 Termos e relacionamentos

O objetivo desta avaliação é verificar se os relacionamentos podem dispensar o uso dos termos como descritores. São avaliadas as estratégias NMRLe, sóRLBe, SINNf, sóSINNf, BIGRf e sóBIGRf:

- NMRLe, com termos nominalizados e RLBs, usa as Equações 7, 8 e 9 para cálculo do peso dos descritores;
- sóRLBe, apenas com RLBs, usa as Equações 7 e 9;
- SINNf, com termos lematizados e sintagmas nominais, usa as Equações 7 e 13;
- sóSINNf, apenas com sintagmas nominais, usa as Equações 7 e 13;
- BIGRf, com termos nominalizados e bigramas, usa as Equações 7 e 13; e
- sóBIGRf, apenas com bigramas, usa as Equações 7 e 13.

Enquanto SINNf, sóSINNf, BIGRf e sóBIGRf são baseadas em freqüência de ocorrência, NMRLe e sóRLBe são baseadas em evidência.

Deve ser salientado que, na estratégia de indexação SINN (utilizada pelas estratégias de busca SINNf e sóSINNf), o espaço de descritores é constituído de 13,3% de sintagmas nominais com apenas um termo. Da mesma forma, a estratégia de indexação BIGR (utilizada pelas estratégias de busca BIGRf e sóBIGRf), inclui em seu espaço de descritores 2,9% de termos, e não bigramas (ver Tabela 4.9). As estratégias sóSINNf e sóBIGRf não utilizam esses termos, mas SINNf e BIGRf sim.

Outro aspecto a considerar é o modo como algumas coincidências, durante o processo de busca, são permitidas em algumas estratégias. Em sóSINNf e em SINNf, um sintagma nominal menor pode coincidir com parte de um sintagma nominal maior. Nas estratégias SINNf e BIGRf, é possível a coincidência de um termo com um componente de um relacionamento (sintagma nominal ou bigrama, respectivamente).

Através da Figura 6.3 e da Tabela 6.4 é possível observar a superioridade dos resultados das estratégias que utilizam termos e relacionamentos sobre as correspondentes estratégias somente com relacionamentos, nos casos de NMRLe e SINNf, respectivamente, sobre sóRLBe e sóSINNf. No caso de SINNf, essa

superioridade se dá especialmente para valores de revocação acima de 0,3.

Por outro lado, somente com valores de revocação acima de 0,55 há superioridade dos resultados de BIGRf sobre sóBIGRf. Isto pode ser observado nas curvas da Figura 6.3, e confirmado através dos dados da Tabela 6.4: os resultados da estratégia sóBIGRf são superiores a BIGRf em revocação, considerando os 10 primeiros documentos recuperados.

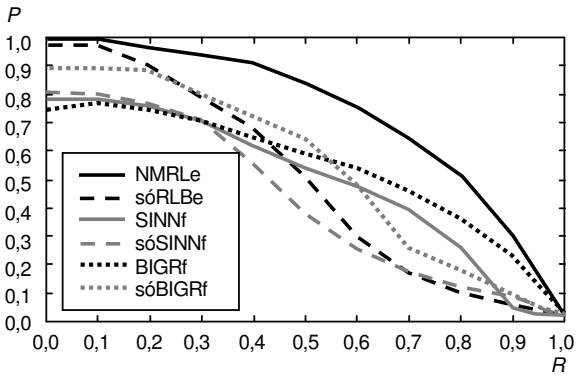

Figura 6.3: Curvas revocação-precisão para estratégias com termos ou com relacionamentos

No caso das RLBs, embora sóRLBe também apresente precisão acima de 0,9, com valores de revocação abaixo de 0,2, essa precisão não chega a superar a de NMRLe.

Tabela 6.4: Alguns resultados para estratégias com termos ou relacionamentos

	precisão		revocação		*F*
estratégias	1	1–10	1–10	1–100	1–10
NMRLe	1,000	0,728	0,659	0,946	0,692
sóBIGRf	0,900	0,616	0,539	0,726	0,575
BIGRf	0,740	0,576	0,529	0,937	0,552
SINNf	0,780	0,546	0,500	0,877	0,522
sóRLBe	0,980	0,560	0,483	0,653	0,519
sóSINNf	0,800	0,500	0,443	0,579	0,470

1 = documento do topo, 1–N = primeiros N documentos

Nas condições desta avaliação, os resultados obtidos demonstram que não é possível dispensar os termos, como descritores, pois apenas os relacionamentos não são suficientes para melhorar a classificação dos documentos por relevância. Isto é verdade no caso de estratégias com sintagmas nominais e RLBs, mas seria inadequado dizer que o mesmo também é válido para estratégias com bigramas. A razão desta diferença pode estar no fato de os bigramas não se comportarem como descritores da mesma forma que os outros tipos de relacionamentos, conforme comentário apresentado na Seção 2.4: em um bigrama, a descrição de um termo é afetada por outro que o antecede. Enquanto em sintagmas nominais e RLBs há informações sintáticas consideradas (diferentes daquelas presentes em descritores constituídos apenas por termos), a captura de um bigrama se dá fundamentalmente através de abordagem estatística. Em outras palavras, quando são acrescentados termos aos descritores constituídos apenas por sintagmas nominais ou RLBs, incluem-se outras informações (estatísticas) à indexação. Nesses casos, há ganho

de descrição e, em decorrência, os resultados da recuperação melhoram (embora não necessariamente quanto à precisão). Por outro lado, quando são acrescentados termos aos descritores constituídos apenas por bigramas, mantém-se a abordagem estatística e não é possível dizer que a descrição melhora.

6.5 Termos, relacionamentos e operadores Booleanos

O objetivo desta avaliação é comparar estratégias que especificam dependência de termos através de RLBs e de consultas Booleanas. São avaliadas as estratégias NMRLeB, NMRLe, NMeB, e NMe:

- NMRLe, com termos nominalizados e RLBs, usa as Equações 7, 8 e 9 para o cálculo do peso dos descritores;
- NMRLeB, idem à anterior com consulta Booleana;
- NMe, apenas com termos nominalizados, usa as Equações 7 e 8; e
- NMeB, idem à anterior com consulta Booleana.

Todas essas estratégias usam cálculo do peso dos descritores baseado em evidência.

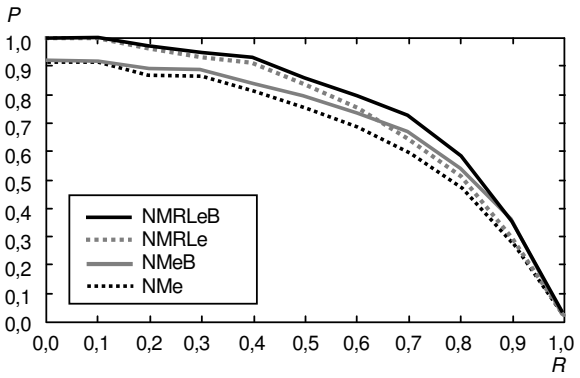

Figura 6.4: Curvas revocação-precisão para alternativas com termos nominalizados, RLBs e operadores Booleanos

Na Figura 6.4, através das curvas revocação-precisão, é possível observar a superioridade das abordagens que usam termos e relacionamentos como descritores (NMRLeB e NMRLe) em relação às demais, que utilizam apenas termos. O mesmo pode ser observado na Tabela 6.5.

Tabela 6.5: Alguns resultados para alternativas com nominalização, RLBs e operadores Booleanos

	precisão		revocação		*F*
	1	1–10	1–10	1–100	1–10
NMRLeB	1,000	0,748	0,690	0,959	0,718
NMRLe	1,000	0,728	0,659	0,946	0,692
NMeB	0,920	0,716	0,666	0,958	0,690
NMe	0,920	0,694	0,632	0,949	0,662

1 = documento do topo, 1–N = primeiros N documentos

Também é possível afirmar que as estratégias que adotam operadores Booleanos

na consulta têm vantagem sobre as que não os adotam: NMRLeB apresenta melhores resultados que NMRLe, o mesmo ocorrendo com NMeB em relação a NMe.

Nas condições desta avaliação, os resultados obtidos demonstram que a inclusão de RLBs, como descritores, e operadores Booleanos, na consulta, melhoram a recuperação. A combinação desses recursos melhora as vantagens obtidas por eles quando usados separadamente: tanto os resultados de NMRLe (com RLBs) quanto os resultados de NMeB (com consulta Booleana) são superiores aos de NMe (sem esses recursos); entretanto, em comparação a todos esses resultados, os de NMRLeB (com ambos os recursos) são os melhores.

6.6 Tamanho da consulta

O objetivo desta avaliação é comparar a influência do tamanho da consulta em dois conjuntos de estratégias com cálculo do peso dos descritores (i) baseado em evidência, com termos e relacionamentos e (ii) baseado em freqüência de ocorrência, somente com termos. São avaliadas as estratégias NMRLeB2, NMRLeB1, NMfB2 e NMfB1:

- NMRLeB2, com termos nominalizados e RLBs, usa as Equações 7, 8 e 9 e consultas com mais de um termo;
- NMRLeB1, idem à anterior, com consultas com apenas um termo;
- NMfB2, apenas com termos nominalizados, usa a Equação 2 e consultas com mais de um termo; e
- NMfB1, idem à anterior, com consultas com apenas um termo.

Todas as estratégias usam operadores Booleanos nas consultas. Foram executadas 17 consultas (do conjunto de 50) com apenas um termo para NMRLeB1 e NMfB1, enquanto que 33 consultas (também do conjunto de 50) com mais de um termo foram executadas para NMRLeB2 e NMfB2.

NMRLeB1 e NMRLeB2 são baseadas em evidência, enquanto NMfB1 e NMfB2 são baseadas em freqüência de ocorrência.

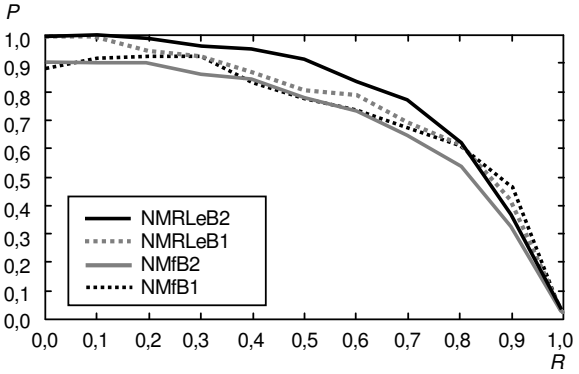

Figura 6.5: Curvas revocação-precisão para estratégias avaliadas através de consultas com tamanhos diferentes

Na Figura 6.5 e na Tabela 6.6 é possível observar que consultas com mais termos favorecem as estratégias baseadas em evidência. A mesma conclusão não é tão clara quanto às estratégias baseadas em freqüência de ocorrência aqui avaliadas. A

diferença entre as curvas revocação-precisão de NMRLeB2 e NMRLeB1, baseadas em evidência, é maior que a diferença entre as curvas revocação-precisão de NMfB2 e NMfB1. Há superioridade dos valores de precisão de NMfB1 sobre NMfB2 com revocação maior que 0,55, mas há alternância entre eles com revocação menor.

Na Tabela 6.6, constatam-se, quanto à medida *F*, diferenças de 0,059 entre NMRLeB2 e NMRLeB1 e de 0,035 entre NMfB2 e NMfB1, ou seja, no primeiro caso a diferença é maior. Além disto, a precisão de NMfB1 é superior à precisão de NMfB2 quando são considerados os 10 primeiros documentos classificados.

Tabela 6.6: Alguns resultados para estratégias em consultas com tamanhos diferentes

	precisão		revocação		*F*
	1	1–10	1–10	1–100	1–10
NMRLeB2	1,000	0,739	0,740	0,961	0,739
NMRLeB1	1,000	0,782	0,602	0,960	0,680
NMfB2	0,909	0.676	0,692	0,967	0,684
NMfB1	0,882	0,747	0,573	0,960	0,649

1 = documento do topo, 1–N = primeiros N documentos

Nas condições desta avaliação, os resultados obtidos demonstram que, em um universo de consultas curtas (com até três termos), as consultas com mais de um termo favorecem a recuperação através das estratégias que usam RLBs e adotam cálculo de peso dos descritores baseado em evidência.

6.7 Cálculo de pesos e operadores Booleanos

O objetivo desta avaliação é analisar estratégias que se baseiam somente na identificação de relacionamentos e no cálculo do peso dos descritores, para tratamento de dependência de termos, em comparação com a proposta apresentada e com outras estratégias que adotam também consulta Booleana. São avaliadas as estratégias NMRLeB, NMRLe, mmNMRLf, SINNfB, SINNf, BIGRfB, btBIGRf e BIGRf:

- NMRLe, com termos nominalizados e RLBs, usa as Equações 7, 8 e 9;
- NMRLeB, idem à anterior, com consulta Booleana;
- mmNMRLf, com termos nominalizados, usa as Equações 2 e 6, conforme a estratégia de Changki Lee e Gary Lee [LEE 2005];
- SINNf, com termos lematizados e sintagmas nominais, usa as Equações 7 e 13;
- SINNfB, idem à anterior, com consulta Booleana;
- BIGRf, com termos nominalizados e bigramas, usa a Equação 7 e 13;
- BIGRfB, idem à anterior, com consulta Booleana; e
- btBIGRf, com bitermos (bigramas sem restrição de ordem dos termos) nominalizados, usa a Equação 5, de acordo com a estratégia de Shrikanth e Srihari [SRI 2002].

Todas as estratégias que usam consultas com operadores Booleanos adotam a mesma formulação para a consulta, conforme foi definida na proposta do modelo TR+.

Além de mmNMRLf, as estratégias que usam bigramas, bitermos e sintagmas nominais são baseadas em freqüência de ocorrência, enquanto NMRLe e NMRLeB são baseadas em evidência. NMRLeB adota o modelo TR+ completo.

Como descritores, mmNMRLf utiliza os termos gerados pela estratégia de indexação NMRL. As RLBs são, também, utilizadas por mmNMRLf, mas apenas para implementar os pares pai-filho necessários a esta estratégia. Por definição, as estratégias com bigramas e bitermos só relacionam termos adjacentes e as estratégias com sintagmas nominais não consideram os verbos ao realizar a indexação dos documentos.

As estratégias avaliadas aqui foram agrupadas em dois gráficos para facilitar a leitura dos mesmos. No primeiro (Figura 6.6) não são incluídas as estratégias com operadores Booleanos na consulta e com sintagmas nominais; no segundo (Figura 6.7) essas alternativas são consideradas.

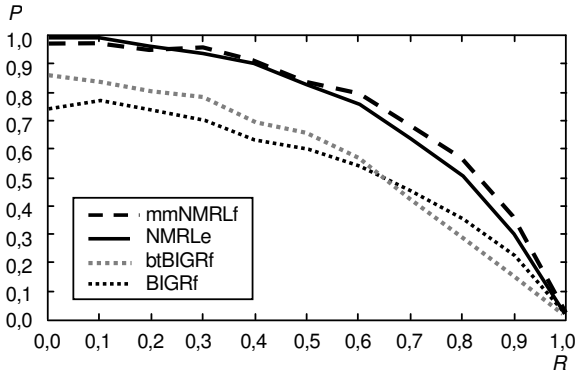

Figura 6.6: Curvas revocação-precisão para alternativas de cálculo de pesos em estratégias sem operadores Booleanos

Na Figura 6.6 é possível observar que nmNMRLf e btBIGRf apresentam resultados superiores, respectivamente, a NMRLe e BIGRf. A causa disto pode estar no fato de as duas primeiras, além de usarem relacionamentos, incluírem formulações específicas para o cálculo do peso dos descritores que trazem vantagens na recuperação dos documentos.

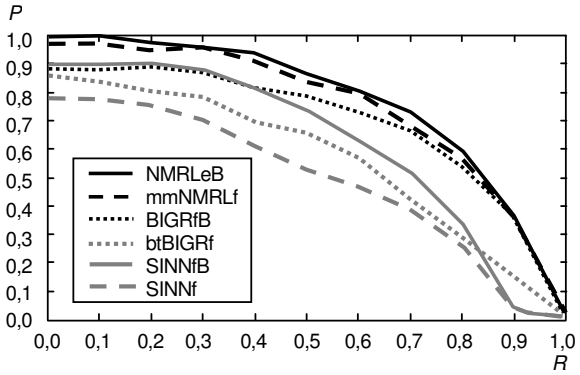

Figura 6.7: Curvas revocação-precisão para alternativas de cálculo de pesos em estratégias com ou sem operadores Booleanos

Na Figura 6.7, entretanto, as estratégias mmNMRLf e btBIGRf têm performances inferiores ao serem comparadas, respectivamente, com NMRLeB e BIGRfB. Nestas duas últimas, a inclusão de operadores Booleanos na consulta faz a

diferença. Também é possível observar, na Figura 6.7, que a estratégia SINNfB apresenta melhores resultados que a SINNf, pela mesma razão.

Tabela 6.7: Alguns resultados para alternativas de cálculo de pesos e uso de operadores Booleanos

	precisão		revocação		*F*
	1	1–10	1–10	1–100	1–10
NMRLeB	1,000	0,748	0,690	0,959	0,718
mmNMRLf	0,960	0,732	0,671	0,961	0,700
NMRLe	1,000	0,728	0,659	0,946	0,692
BIGRfB	0,880	0,704	0,660	0,964	0,681
SINNfB	0,900	0,650	0,599	0,878	0,623
btBIGRf	0,860	0,612	0,560	0,912	0,585
BIGRf	0,740	0,576	0,529	0,937	0,552
SINNf	0,780	0,546	0,500	0,877	0,522

1 = documento do topo, 1–N = primeiros N documentos

Os dados da Tabela 6.7 mostram que as RLBs trazem vantagens para as estratégias que as usam, tanto como descritores (em NMRLeB e NMRLe), quanto apenas como apoio ao cálculo do peso dos descritores (em mmNMRLf). Levando em conta, ainda, os tipos de relacionamentos (sintagmas nominais, RLBs e bigramas ou bitermos), sem o uso de consulta Booleana, os piores resultados são obtidos pelos sintagmas nominais, ficando os bigramas (e os bitermos) com performance intermediária. Na Figura 6.7, é possível verificar que, com consulta Booleana, com valores acima de 0,4 de revocação, os sintagmas nominais apresentam valores menores de precisão quando comparados com os bigramas. A situação se inverte para valores de revocação abaixo de 0,4. SINNfB apresenta maior precisão que BIGFfB, considerando os documentos do topo da classificação (ver Tabela 6.7). Uma provável razão para este fato é que a estratégia com sintagmas nominais não inclui os verbos entre seus descritores e, assim, recupera um número menor de documentos relevantes, no total, quando comparada com a estratégia com bigramas.

Nas condições desta avaliação, os resultados obtidos demonstram que a inclusão de operadores Booleanos na consulta complementa o cálculo da representatividade dos descritores, melhorando a classificação dos documentos por relevância. Também demonstram a superioridade, quanto aos resultados da recuperação, da estratégia com nominalização e RLBs, seguindo o modelo TR+ completo, sobre as demais estratégias.

6.8 Resumo do Capítulo

Nas condições das avaliações realizadas, os resultados obtidos demonstram que uma estratégia melhora sua performance, quanto à classificação dos documentos por relevância, se:

- utiliza a nominalização em alternativa à lematização e ao *stemming*, como procedimento de normalização lexical;
- não dispensa os termos, como descritores, e utiliza relacionamentos, especialmente RLBs, para tratar dependência de termos;
- utiliza cálculo do peso dos descritores baseado em evidência, em substituição ao cálculo baseado simplesmente em freqüência de ocorrência; e
- combina relacionamentos, como descritores, e operadores Booleanos na

consulta, para especificar dependência de termos.

Os resultados também demonstram que, em um universo de consultas curtas (com até três termos), as consultas com mais de um termo favorecem as estratégias com RLBs, ou seja, melhoram seus resultados de recuperação.

Na Tabela 6.8 são apresentados alguns resultados das principais estratégias comparadas através das avaliações realizadas.

Tabela 6.8: Alguns resultados das principais estratégias avaliadas

	precisão		revocação		*F*
	1	1–10	1–10	1–100	1–10
NMRLeB	1,000	0,748	0,690	0,959	0,718
mmNMRLf	0,960	0,732	0,671	0,961	0,700
NMRLe	1,000	0,728	0,659	0,946	0,692
BIGRfB	0,880	0,704	0,660	0,964	0,681
NMf	0,900	0,692	0,636	0,953	0,663
SINNfB	0,900	0,650	0,599	0,878	0,623
LMf	0,920	0,640	0,583	0,881	0,610
STf	0,900	0,626	0,587	0,934	0,606
btBIGRf	0,860	0,612	0,560	0,912	0,585
VRf	0,920	0,600	0,529	0,808	0,562
BIGRf	0,740	0,576	0,529	0,937	0,552
SINNf	0,780	0,546	0,500	0,877	0,522
sóRLBe	0,980	0,560	0,483	0,653	0,519

1 = documento do topo, 1–N = primeiros N documentos

As relações custo/beneficio das estratégias apresentadas na Tabela 6.8 são analisadas no próximo Capítulo, assim como a análise de complexidade dos algoritmos utilizados.

7 ANÁLISE DE COMPLEXIDADE E DA RELAÇÃO CUSTO/BENEFÍCIO

7.1 Introdução

Neste Capítulo é apresentada a análise de complexidade dos algoritmos utilizados nas estratégias avaliadas no Capítulo anterior e são examinadas as suas relações custo/benefício. O custo de cada estratégia é considerado através:

- do tamanho do espaço de descritores, considerando a memória necessária para armazenar as informações decorrentes da fase de indexação e
- do tempo de processamento necessário para construir o espaço de descritores, durante a fase de indexação, e para atender a consulta, durante a fase de busca.

Os dados sobre os custos das estratégias examinadas são relacionados à performance de cada uma dessas estratégias para classificar os documentos por relevância em atendimento às consultas. Desta forma é determinada a relação custo/benefício de cada estratégia.

7.2 Tamanho do espaço de descritores

Todas as estratégias de indexação examinadas são incluídas nesta análise. Elas são resumidas a seguir:

- VR, cujos termos são variantes das palavras conforme ocorrem no texto dos documentos;
- LM, com termos lematizados;
- ST, com termos na forma de *stems*;
- NM, com termos nominalizados;
- BIGR, com bigramas com termos nominalizados;
- SINN, com sintagmas nominais com termos lematizados; e
- NMRL, com termos nominalizados e RLBs.

Na Tabela 7.1 são apresentados tanto o espaço de memória gasto pelos arquivos de índice de cada estratégia quanto a quantidade de descritores de cada uma.

Na estratégia BIGR, 2,9% dos descritores são, na verdade, termos que ocorrem de forma isolada no texto (ver Tabela 4.9). O mesmo acontece com a estratégia SINN: 13,3% dos sintagmas nominais possuem apenas um termo e são tratados também como sintagmas nominais. Todos os descritores de BIGR foram incluídos em um único arquivo de índice. A mesma decisão foi tomada em relação à estratégia SINN. Apenas NMRL utiliza um arquivo de índice específico para os termos e outros três para os relacionamentos (RLBs do tipo classificação, restrição e associação), conforme é apresentado na Tabela 7.1.

Todos os arquivos de índice foram implementados na forma de listas invertidas. Trechos desses arquivos são apresentados, como exemplo, no Anexo C.

A estratégia mais econômica (em quantidade de descritores e, também, em tamanho do arquivo de índice) é a ST (ver Tabela 7.1). Isto acontece porque o *stemming* é mais impactante, quanto à redução de espaço, em comparação com os outros processos de normalização lexical. Ele produz um número menor de termos diferentes normalizados e cada um deles possui, em média, menor número de caracteres.

Tabela 7.1: Informações sobre os arquivos de índice

Estratégia de indexação	arquivo invertido (Kb)	descritores		
		termos	relaciona-mentos	total
VR	3.673	57.366	0	57.366
LM	3.222	37.267	0	37.267
ST	2.873	24.013	0	24.013
NM	3.713	36.479	0	36.479
SINN	4.453	0	112.239	112.239
BIGR	8.938	0	296.243	296.243
NMRL (total)	10.542	38.616	245.093	283.709
NMRL (termos)	4.557	38.616		
NMRL (classificações)	1.366		54.714	
NMRL (associações)	909		29.613	
NMRL (restrições)	3.710		160.766	

As quantidades de termos em LM e NM são semelhantes, ainda que NM, algumas vezes, indexe dois termos (os substantivos abstrato e concreto) derivados de uma única palavra nominalizada. Essas duplicações são, no total, compensadas pelo menor número de termos diferentes derivados na nominalização, quando comparada com a lematização. Por exemplo, a partir dos adjetivos "preferencial" e "preferível" é derivado um único substantivo ("preferência") na nominalização, enquanto que as formas lematizadas são duas, no caso, iguais às originais.

Ainda que NM tenha menor número de termos que VR, o arquivo de índice de NM é 2% maior que o da estratégia VR. Tal situação ocorre porque o tamanho médio dos termos nominalizados (principalmente dos substantivos abstratos) é maior que o tamanho médio das formas originais das respectivas palavras.

As quantidades de termos das estratégias NM e NMRL são diferentes porque NMRL inclui, na indexação, "termos negados" (quando relacionados a advérbios ou à conjunção "nem") como termos diferentes dos termos originais.

Entre as estratégias que utilizam relacionamentos, SINN é a mais econômica em espaço de memória e, também, em número de descritores. As estratégias menos econômicas são a NMRL, em espaço de memória, e a BIGR, em número de descritores.

7.2.1 Evolução de crescimento do espaço de descritores

Quanto ao espaço de memória e ao número de descritores, há preocupação em estimar o crescimento dos arquivos de índice à medida que novos documentos são incorporados à coleção e indexados.

A Figura 7.1 apresenta a quantidade de descritores relacionada à quantidade de itens de texto dos documentos indexados em cada uma das estratégias.

A Figura 7.2 apresenta a relação entre o tamanho dos arquivos de índice e a quantidade de itens de texto dos documentos indexados.

Os valores apresentados foram obtidos através da indexação da coleção

fracionada a cada 100.000 itens de texto.

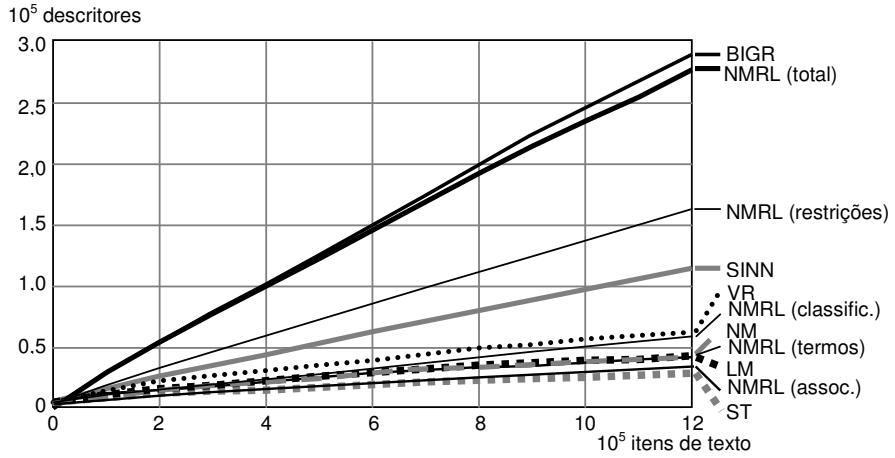

Figura 7.1: Relação entre descritores e itens de texto nos documentos

É usual que seja observada uma tendência de suavização das curvas de crescimento do espaço de descritores à medida que os documentos vão sendo inseridos nos índices [LAH 2000]. Isto ocorre porque, a cada novo documento indexado, diminui a probabilidade de aparecimento de novos descritores, diferentes dos já incluídos nos índices. Tal tendência de suavização é, entretanto, mínima, no experimento aqui realizado.

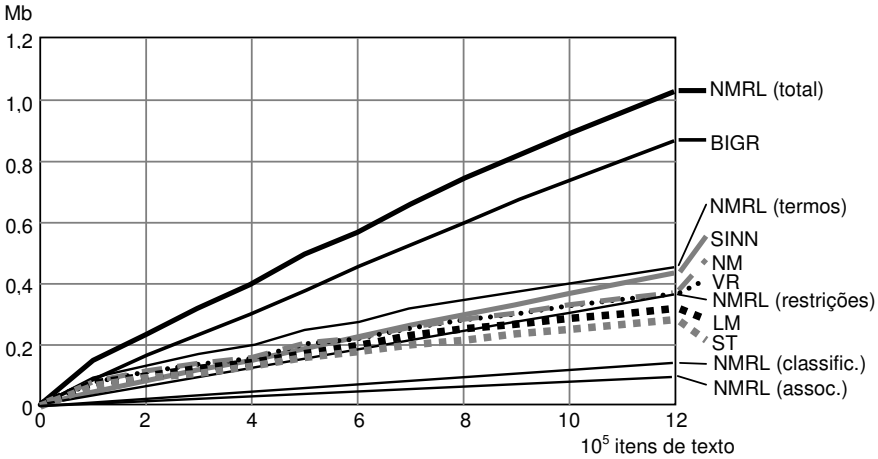

Figura 7.2: Relação entre tamanho dos arquivos de índice e itens de texto nos documentos

Patamares de horizontalização das curvas de crescimento podem ser estimados através de alguns limites encontrados no léxico, para o Português:

- a quantidade de substantivos: pouco mais de 80.000 substantivos, e
- a quantidade de adjetivos, advérbios, substantivos, particípios e verbos: pouco mais de 140.000 palavras.

O número de substantivos estabelece um patamar para a quantidade de termos, para as estratégias NM e NMRL. Este patamar só deve ser ultrapassado em razão da ocorrência de nomes próprios e, no caso de NMRL, de "termos negados".

O segundo limite (a quantidade de adjetivos, advérbios, substantivos, particípios e verbos) influencia o patamar para a estratégia LM.

Quanto à estratégia ST, pode-se esperar que a quantidade de *stems* fique abaixo do patamar da estratégia LM, se mantida a tendência atual, ou seja, quantidade de *stems* menor que a de lemas nos índices. Por outro lado, a quantidade de variantes para os termos que constituem a estratégia VR deve ser maior que LM devido à quantidade de flexões diferentes das palavras, característica da língua Portuguesa, rica em desinências nominais e verbais [KEH 2000].

Os descritores do tipo relacionamento (em SINN, BIGR e NMRL) devem ter a suavização de suas curvas, quando comparadas às dos termos, em coordenadas com valor maior de abscissa, ou seja, com maiores quantidades de itens de texto. Isto acontece porque os descritores dessas estratégias são gerados por combinação de termos. Entretanto, a suavidade das curvas também acontece pois, assim como ocorre com os termos, há relacionamentos de uso mais comum, que são logo inseridos na indexação, e outros menos comuns, que aparecem mais raramente e contribuem pouco para o crescimento dos respectivos arquivos de índice.

Essas estimativas sobre a suavização das curvas de crescimento dos espaços de descritores necessitam confirmação através da indexação de coleção de documentos maior que a utilizada nos experimentos aqui realizados, ou seja: com mais documentos e, principalmente, maior quantidade de itens de texto.

7.3 Algoritmos e análise de complexidade

Nesta Seção são apresentados os algoritmos das fases de indexação e de busca, implementados para as estratégias avaliadas, e é realizada a análise de complexidade dos mesmos, sempre considerando o pior caso quanto ao tempo de processamento. A análise desenvolvida adota os fundamentos apresentados por Rawlins [RAW 92].

7.3.1 Algoritmos de cada fase

Nesta Seção, os algoritmos analisados a seguir quanto à complexidade são relacionados.

Os algoritmos da fase de indexação são:

- ***Toquenização***: os itens de texto são identificados.
- **Etiquetagem**: os itens de texto são marcados com as correspondentes etiquetas morfológicas.
- **Normalização**: as formas normalizadas das palavras originais são derivadas através de processos de lematização, *stemming* ou nominalização, conforme cada estratégia.
- **Geração de descritores de um documento**: todos os descritores, descartadas as duplicatas, são listados, cada um com os valores necessários para o cálculo do respectivo peso, conforme cada estratégia.
- **Inclusão no índice**: os dados da listagem anterior são incluídos no arquivo de índice correspondente, conforme cada estratégia.

Os algoritmos da fase de busca são:

- **Formulação da consulta**[19]: inclui ***toquenização***, **etiquetagem** e **normalização**, como na fase de indexação, e, ainda, **preparação da consulta**. Na preparação, termos e, quando é o caso, relacionamentos da consulta são identificados e, conforme a estratégia, associados aos respectivos pesos. Se a estratégia utiliza operadores Booleanos, eles são incluídos.
- **Pesquisa**: termos (e relacionamentos, quando é o caso) da consulta são pesquisados nos arquivos de índice, buscando coincidência com os descritores dos documentos. São obtidas listagens com os descritores coincidentes, cada um com a relação dos documentos onde ocorre e os dados necessários para o cálculo do valor de relevância desses documentos.
- **Classificação**: o valor de relevância de cada documento recuperado é calculado de acordo com os dados obtidos no procedimento anterior. Os documentos recuperados são classificados por ordem decrescente do valor de relevância de cada um, conforme os critérios de cada estratégia.

Os algoritmos para *toquenização*, etiquetagem morfológica e lematização analisados quanto à complexidade são os mesmos que foram testados quanto ao tempo de processamento (Seção 7.4). Eles foram desenvolvidos em trabalhos de diplomação por alunos de graduação [TOS 2002, RAZ 2003]. Esses algoritmos não são os mesmos cujos resultados foram avaliados no Capítulo 5. Naquela avaliação foram consideradas as informações constantes da coleção de documentos utilizada, onde os procedimentos para *toquenização*, etiquetagem e lematização já haviam sido realizados. Quanto a esses algoritmos, o que se pretende aqui é apresentar a analise da complexidade de alternativas para esses procedimentos.

7.3.2 *Toquenização* e etiquetagem

O algoritmo para *toquenização*, apresentado na Figura 7.3, tem complexidade:

$$O(37|C|3+|C|) = O(|C|),$$

sendo |C| a quantidade de caracteres do texto.

```
Para cada c ∈ C = {caracteres do texto}
   Para cada cE do conjunto de 37 caracteres especiais (espaço,
   pontuações e outros)
      Se c=cE então
         Identificar nova palavra do texto
         Se cE é pontuação então identificar pontuação
         Inicializar palavra e interromper comparação com atual c
   Se não houve sucesso na comparação então
      Acumular c para formação de nova palavra
```

Figura 7.3: Algoritmo para *toquenização*

Logo, o algoritmo para *toquenização* apresenta complexidade da classe linear $O(n)$.

[19] Não confundir com a etapa de formulação da consulta pelo usuário, que depende da interface disponibilizada pelo sistema de RI utilizado e é um procedimento manual. O algoritmo para formulação da consulta, referido aqui, corresponde ao processamento automático dos dados informados pelo usuário (uma lista de termos ou uma frase) para gerar a entrada do algoritmo de Pesquisa.

```
Para cada p ∈ P = {palavras do texto}
   Para cada pL ∈ PL = {palavras de mini-léxico}
      Se p=pL então etiquetar p e interromper pesquisa em PL
Para cada p ∈ PET = {palavras do texto por etiquetar}
   Pesquisar sufixo em árvore ternária de pesquisa com NET nodos
      Se encontra sufixo então etiquetar p e interromper pesquisa
Para cada p ∈ P
   Para cada n ∈ L = {núcleos em base de núcleos de locuções}
      Se p=n então
         Verificar 1 palavra anterior e 1 posterior a p
         Se verificação tem sucesso então
            Identificar locução e interromper pesquisa em L
Para cada p ∈ PET
   Para cada r ∈ RSX = {regras sintáticas}
      Se r é aplicável a p e 2 palavras anteriores e 2
      posteriores então
         Etiquetar p e interromper aplicação de regras
```

Figura 7.4: Algoritmo para etiquetagem de texto

O algoritmo para etiquetagem, apresentado na Figura 7.4, tem complexidade:

$$O(|P||P_L|)+O(|P_{ET}|\log_3 N_{ET})+O(3|P||L|)+O(5|P_{ET}||R_{SX}|). \quad (14)$$

Como $|P| \geq |P_{ET}|$, substituindo em (14) é obtida a complexidade:

$$O(|P|(|P_L|+\log_3 N_{ET}+|L|+|R_{SX}|)). \quad (15)$$

Considerando a constante $k_{ET} = |P_L|+N_{ET}+|L|+|R_{SX}| \geq |P_L|+\log_3 N_{ET}+|L|+|R_{SX}|$ (já que a árvore está parcialmente balanceada), substituindo em (15) é obtida a complexidade:

$$O(k_{ET}|P|) = O(|P|),$$

sendo $|P|$ a quantidade de palavras do texto.

Logo, o algoritmo para etiquetagem morfológica do texto apresenta complexidade da classe linear $O(n)$.

7.3.3 Normalização lexical

Nesta Seção são analisados algoritmos para lematização, para *stemming* e para nominalização.

```
Para cada p ∈ P = {palavras do texto}
   Verificar etiqueta morfológica de p
   Se p é adjetivo, artigo, numeral, pronome, substantivo ou
   verbo então
      Pesquisar sufixo em árvore ternária de pesquisa
      /* Há uma árvore para cada categoria mencionada de p, sendo
      que a maior tem NLM nodos */
         Se encontra sufixo então
            Derivar lema de p e interromper pesquisa em árvore
   Se não foi derivado lema para p então lema de p é p
```

Figura 7.5: Algoritmo para lematização

O algoritmo para lematização, apresentado na Figura 7.5, tem complexidade:

$$O(|P|(1+\log_3 N_{LM}+1)). \quad (16)$$

Considerando a constante $k_{LM} = 1+N_{LM}+1 \geq 1+\log_3 N_{LM}+1$ (já que a árvore está parcialmente balanceada), substituindo em (16) é obtida a complexidade:

$$O(|P|k_{LM}) = O(|P|),$$

sendo |P| a quantidade de palavras do texto.

```
Para cada p ∈ P = {palavras do texto}
   Para cada pE ∈ PE = {palavras da base de exceções}
      Se p=pE então derivar stem de p e interromper pesquisa em PE
Para cada p ∈ PST = {palavras do texto ainda sem stem}
   Pesquisar sufixo em árvore ternária de pesquisa com NST nodos
      Se encontra sufixo então
         Derivar stem de p e interromper pesquisa na árvore
   Se sufixo de p não é encontrado então stem de p é p
```

Figura 7.6: Algoritmo para *stemming*

O algoritmo para *stemming*, apresentado na Figura 7.6, tem complexidade:

$$O(|P||P_E|)+O(|P_{ST}|(\log_3 N_{ST}+1)). \quad (17)$$

Considerando as constantes $k_{EX} = |P_E|$ e $k_{ST} = N_{ST} \geq \log_3 N_{ST}$ (já que a árvore está parcialmente balanceada), e como $|P| \geq |P_{ST}|$, substituindo em (17) é obtida a complexidade:

$$O(|P|(k_{EX}+k_{ST})) = O(|P|),$$

sendo |P| a quantidade de palavras do texto.

```
Para cada p ∈ P = {palavras do texto}
   Verificar etiqueta morfológica de p
   Se p é advérbio então transformar em adjetivo correspondente
   Se p é adjetivo ou verbo então
      Pesquisar em autômato de exceções implementado em árvore
      ternária de pesquisa com NE nodos
      Se pesquisa tem sucesso então derivar substantivos de p
      Senão
         Pesquisar p em autômato de adjetivos ou de verbos
         implementado em árvore ternária de pesquisa com NA (para
         adjetivos) ou NV (para verbos) nodos
         Se pesquisa tem sucesso então derivar substantivos de p
         Senão não há nominalização para p
   Senão se p é substantivo
      Pesquisar em autômato de sinônimos implementado em árvore
      ternária de pesquisa com NS nodos
      Se pesquisa tem sucesso então derivar sinônimo de p
      Senão não há sinônimo de p
```

Figura 7.7: Algoritmo para nominalização

O algoritmo para nominalização é apresentado na Figura 7.7. O detalhamento da pesquisa nos autômatos é apresentado no Anexo A, na Figura A.1. Assumindo $N_M=\max(N_A,N_V,N_S)$, ou seja, sendo N_M o maior número de nodos considerando as árvores ternárias utilizadas, esta pesquisa apresenta complexidade:

$$O(1+1+\log_3 N_E+\log_3 N_M),$$

ou, no caso dos substantivos,

$$O(1+1+\log_3 N_M).$$

Considerando a constante $k_{NM} = 2+N_E+N_M \geq 2+\log_3 N_E+\log_3 N_M \geq 2+\log_3 N_M$ (já que as árvores estão parcialmente balanceadas) e que o número de adjetivos, advérbios, verbos ou substantivos é, no máximo, igual a |P|, o algoritmo para nominalização apresenta a complexidade:

$$O(|P|k_{NM}) = O(|P|),$$

sendo |P| a quantidade de palavras do texto.

Logo, os algoritmos para normalização lexical (lematização, *stemming* e nominalização) apresentam complexidade da classe linear $O(n)$.

7.3.4 Geração de descritores de um documento

Nesta Seção são analisados os algoritmos para geração de descritores de um documento no caso de estratégias com unigramas ou bigramas, de estratégias com sintagmas nominais e de estratégias com termos nominalizados e RLBs.

No algoritmo da Figura 7.8, os descritores correspondem a termos, na estratégia com unigramas, ou a bigramas, na estratégia que usam esses últimos descritores. Este algoritmo para geração de descritores de um documento apresenta a complexidade:

$$O(|P|(|L_d|+1)). \qquad (18)$$

Assumindo, como pior caso, que o número de termos seja igual ao de palavras do texto, ou seja, |T| = |P|, e sendo $|P| \geq |L_d|$, substituindo em (18) é obtida a complexidade:

$$O(|T|^2+|T|) = O(|T|^2),$$

sendo |T| a quantidade de termos.

No caso dos bigramas, substituindo T por *I*, a complexidade é:

$$O(|I|^2),$$

sendo |*I*| a quantidade de bigramas.

```
Para cada p ∈ P = {palavras do texto}
   Verificar etiqueta morfológica de p
   Se p não é stopword então
      Ler descritor d (correspondente a p) ∈ D = {descritores do
      texto, descartadas as stopwords}
      Para cada dL ∈ Ld = {descritores já listados}
         Se d = dL então
            Incrementar valor de freqüência de dL
            Interromper pesquisa em Ld
         Senão Se d < dL (alfabeticamente) então
            Incluir d antes de dL (com freqüência 1) em Ld
            Interromper pesquisa em Ld
      Se d ∉ Ld então incluir d ao final de Ld (com freqüência 1)
```

Figura 7.8: Algoritmo para geração de descritores de um documento para estratégias com unigramas ou bigramas

Os algoritmos para geração de descritores de um documento para estratégias com sintagmas nominais (Figura 7.10) e para identificar RLBs (Figura 7.11) usam o algoritmo para identificação de frases e componentes de frases do texto (Figura 7.9).

Este algoritmo apresenta complexidade:

$$O(|P|2+|F||P_F|). \quad (19)$$

Como $|P| = |F||P_F|$ em um texto com $|P|$ palavras, substituindo em (19) é obtida a complexidade:

$$O(|P|2+|P|) = O(|P|). \quad (20)$$

```
Para cada p ∈ P = {palavras do texto}
   Se p é conjunção ou pronome relativo ou há pontuação após p
   então
      Identificar nova frase do texto
      Inicializar frase para nova identificação
   Senão acrescentar p em frase em formação
Para cada f ∈ F = {frases do texto}
   Para cada pF ∈ PF = {palavras de f}
      Identificar delimitador entre lado esquerdo e conjunto
      verbal e entre conjunto verbal e lado direito
```

Figura 7.9: Algoritmo para identificação de frases e componentes de frases

Considerando (20), o algoritmo para geração de descritores de um documento para estratégias com sintagmas nominais (Figura 7.10) apresenta a complexidade:

$$O(|P|)+O(|F|2(|P_C|+1))+O(|S|(|L_S|+1)). \quad (21)$$

Como $|P| \geq |F|2|P_C| \geq |F|2$, $|P| \geq |S|$ e $|P| \geq |L_S|$ em um texto com $|P|$ palavras, substituindo em (21) é obtida a complexidade:

$$O(|P|+2|P|+|P|^2+|P|) = O(4|P|+|P|^2) = O(|P|^2),$$

sendo $|P|$ a quantidade de palavras do texto.

```
Identificar frases e componentes de frases do texto
Para cada f ∈ F = {frases do texto}
   Para cada componente (lados esquerdo e direito) de f
      Para cada p ∈ PC = {palavras do componente de f}
         Se é palavra inicial e final e não é substantivo,
         adjetivo ou particípio então descartar p
      Identificar novo sintagma nominal
Para cada s ∈ S = {sintagmas nominais do texto}
   Para cada sL ∈ LS = {sintagmas nominais já listados}
      Se s = sL então
         Incrementar valor de freqüência de sL
         Interromper pesquisa em LS
      Senão se s < sL (alfabeticamente) então
         Incluir s (com freqüência 1) em LS e interromper pesquisa
   Se s ∉ LS então Incluir s ao final de LS (com freqüência 1)
```

Figura 7.10: Algoritmo para geração de descritores de um documento para estratégias com sintagmas nominais

O algoritmo para identificação de RLBs de um documento (Figura 7.11) utiliza, para identificar frases e componentes de frases, o algoritmo da Figura 7.9, considerando sentenças com, no máximo, $|Pg|$ palavras. Assim, o algoritmo da Figura 7.11 apresenta a

complexidade:

$$O(|G|(|Pg|+|Fg|2+|Fg|+|Fg|3(|P_C|(|P_C|-1)|R_{CR}|+|R_{AS}|))+|Fg|3|P_C|)). \quad (22)$$

Como $|P| = |G||Pg| = |G||Fg|3|P_C| > |G||Fg|3$ e $|P| > |P_C|-1$ em um texto com $|P|$ palavras e $|R|$ RLBs, e como $|R_{CR}| = 21$ e $|R_{AS}| = 3$, substituindo em (22) é obtida a complexidade:

$$O(21|P|^2+5|P|+|P|) = O(|P|^2).$$

```
Para cada g ∈ G = {sentenças do texto}
   Identificar frases e componentes de cada frase
   Para cada f ∈ Fg = {frases de g}
      Estruturar f em g
      Identificar núcleos de f
   Para cada f ∈ Fg propagar núcleos de f
   Para cada f ∈ Fg
      Para cada componente de f
         Para cada p1 ∈ PC = {palavras do componente de f}
            Se p1 é adjetivo, advérbio, substantivo ou verbo então
               Para cada p2 ∈ PC-{p1}
                  Se p2 é adjetivo, advérbio, substantivo ou verbo
                  então
                     Para cada r ∈ RCR = {regras para classificação
                     ou restrição}
                        Se r é aplicavel à dependência entre p1 e p2
                        então
                           Capturar nova RLB e interromper aplicação
         Se há núcleo no componente de f
            Para cada r ∈ RAS = {regras para associação}
               Se r é aplicável à dependência entre núcleos então
                  Capturar nova RLB e interromper aplicação
   Para cada núcleo de componente de frase de g
      Propagar RLB que relaciona o núcleo com alguma outra
      palavra do componente
```

Figura 7.11: Algoritmo para identificar RLBs de um documento

O algoritmo da Figura 7.12 utiliza, para gerar termos nominalizados, o algoritmo da Figura 7.8 e, para identificar RLBs, o algoritmo da Figura 7.11. Assim, o algoritmo para geração de descritores de um documento para estratégias com termos nominalizados e RLBs apresenta a complexidade:

$$O(|T|^2)+O(|P|^2)+O(|R|(|L_R|+1))+O(|T|(|L_R|+1))+O(|L_R|(|T|+1)). \quad (23)$$

Assumindo, como pior caso, que o número de termos seja igual ao de palavras do texto, ou seja, $|T| = |P|$, e como $|R| \geq |L_R|$, substituindo em (23) é obtida a complexidade:

$$O(|T|^2+|T|^2+|R|(|R|+1)+|T|(|R|+1)+|R|(|T|+1)) =$$
$$O(2|T|^2+|R|^2+|R|+|T||R|+|T|+|R||T|+|R|) =$$
$$O(3|T|^2+2|R|+|R|^2+|T||R|+|R||T|). \quad (24)$$

Como $|I| \geq |T|$ e $|I| \geq |R|$, já que $I = T \cup R$, substituindo em (24), é obtida a complexidade:

$$O(3|I|^2+3|I|+|I|^2+|I|^2+|I|^2) = O(6|I|^2+3|I|) = O(|I|^2),$$

sendo $|I|$ a quantidade de descritores (termos e relacionamentos).

Logo, os algoritmos para geração de descritores de um documento, nas

estratégias examinadas, apresentam complexidade da classe quadrática $O(n^2)$.

```
Gerar termos nominalizados
Capturar RLBs
Para cada r ∈ R = {RLBs do texto}
   Para cada r_L em L_R = {RLBs já listadas}
      Se r = r_L então
         Incrementar valor de freqüência e interromper pesquisa
      Senão r < r_L (alfabeticamente) então
         Incluir r antes de r_L(com freqüência 1) em L_R
         Interromper pesquisa em L_R
   Se r ∉ L_R então
      Incluir r ao final de L_R (com freqüência 1)
Para cada t ∈ T = {termos}
   Para cada r ∈ L_R
      Se t é argumento de r então
         Incrementar valor de participação em mecanismo de coesão
   Calcular evidência de t
Para cada r ∈ L_R
   Para cada t ∈ T
      Se t é argumento de r então
         Armazenar valor de evidência de t para r
   Calcular evidência de r
```

Figura 7.12: Algoritmo para geração de descritores de um documento para estratégias com termos nominalizados e RLBs

7.3.5 Inclusão no índice

No procedimento de inclusão dos descritores de um documento no índice, esses descritores são incluídos no arquivo de índice correspondente, conforme cada estratégia.

```
Para cada t_d ∈ L_d = {descritores capturados de um documento}
   Para cada t ∈ T = {descritores já indexados}
      Se t_d = t então
         Incluir (na lista de documentos de t) a referência ao
         documento e a freqüência de t_d
         Recalcular IDF para t /* se a estratégia exige */
         Interromper pesquisa em T
      Senão se t_d < t (alfabeticamente) então
         Incluir t_d (antes de t) com a referência ao documento e a
         freqüência de t_d
         Calcular IDF para t_d /* se a estratégia exige */
         Interromper pesquisa em T
    Se t_d ∉ T então incluir ao final de T com  referência ao
    documento e freqüencia de t_d
```

Figura 7.13: Algoritmo para inclusão de termos (ou bigramas ou sintagmas nominais) no índice

O algoritmo para inclusão de termos no índice (Figura 7.13) apresenta complexidade:

$$O(|L_d||T|2). \quad (25)$$

Como T tende a ser maior que L_d, substituindo em (25) é obtida a complexidade:

$$O(2|T|^2) = O(|T|^2),$$

sendo |T| a quantidade de termos.

O algoritmo da Figura 7.13 é o mesmo para a inclusão de bigramas ou sintagmas nominais. Portanto, a complexidade para esses algoritmos é a mesma, substituindo-se T por *I*:

$$O(|I|^2),$$

sendo $|I|$ a quantidade de descritores (bigramas ou sintagmas nominais).

A inclusão dos termos nominalizados também é realizada de acordo com o algoritmo apresentado na Figura 7.13, substituindo-se "freqüência" por "evidência". Deve ser considerado, nesse caso, que podem ser derivados até dois substantivos por palavra do texto. Portanto, a análise deve assumir, para os termos nominalizados, que $2|T| \geq |L_d|$. Ainda assim, substituindo em (25) é obtida a mesma complexidade:

$$O(2|T||T|2) = O(4|T|^2) = O(|T|^2),$$

sendo |T| a quantidade de termos nominalizados.

```
Incluir termos nominalizados
Para cada r_d ∈ L_R = {RLBs capturadas de um documento}
   Para cada r ∈ R = {RLBs} /* já indexadas */
      /* r_d = id_d(arg1_d,arg2_d) e r = id(arg1,arg2) */
      Se id_d = id então
         Se arg1_d = arg1 então
            Se arg2_d = arg2 então
               Incluir a referência ao documento e o peso de r_d (na
               lista de documentos de arg2 de r) em R
               Interromper pesquisa em R
            Senão se arg2_d < arg2 (alfabeticamente) então
               Incluir arg2_d (antes de arg2) com a referência ao
               documento e o peso de r_d em R
               Interromper pesquisa em R
         Senão se arg1_d < arg1 (alfabeticamente) então
            Incluir arg1_d (antes de arg1) seguido de arg2_d com a
            referência ao documento e peso de r_d em R
            Interromper pesquisa em R
      Senão se id_d < id (alfabeticamente) então
         Incluir id_d (antes de id) seguido de arg1_d e de arg2_d,
         este com referência ao documento e peso de r_d em R
         Interromper pesquisa em R
   Se r_d ∉ R então
      Incluir id_d (ao final de R) seguido de ar g1_d e de arg2_d,
      este com referência ao documento e peso de r_d em R
```

Figura 7.14: Algoritmo para inclusão de termos nominalizados e RLBs no índice

O algoritmo para inclusão de termos nominalizados e RLBs no índice (Figura 7.14) apresenta a complexidade:

$$O(|T|^2)+O(|L_R||R|2). \quad (26)$$

Considerando que R tende a ser maior que L_R e como $|I| \geq |T|$ e $|I| \geq |R|$, já que $I = T \cup R$, substituindo em (26) é obtida a complexidade:

$$O(3|I|^2) = O(|I|^2),$$

sendo $|I|$ a quantidade de descritores (termos e relacionamentos).

Logo, nas estratégias examinadas, os algoritmos para inclusão de descritores de um documento no índice apresentam complexidade da classe quadrática $O(n^2)$.

7.3.6 Formulação da consulta

O algoritmo para formulação da consulta é apresentado na Figura 7.15. Os mesmos algoritmos desenvolvidos para os documentos são aplicados à consulta. "Obter termos e/ou relacionamentos e seus pesos", realizado através dos mesmos algoritmos para geração de descritores (Seção 7.3.4), e "Incluir operadores Booleanos" constituem a preparação da consulta. O algoritmo para formulação da consulta, então, não considerando a inclusão de operadores Booleanos, apresenta a complexidade:

$$O(|C|)+O(|P|)+O(|P|)+O(|I|^2), \tag{27}$$

para uma consulta com $|C|$ caracteres, $|P|$ palavras e $|I|$ descritores. No caso dos sintagmas nominais $O(|I|^2)$ deve ser substituído por $O(|P|^2)$.

```
Tokenizar
Etiquetar
Normalizar
Obter termos e/ou relacionamentos e seus pesos
Incluir operadores Booleanos /*se a estratégia exige */
```

Figura 7.15: Algoritmo para formulação da consulta

Considerando consultas curtas como as utilizadas, $|P| \leq 3$. Pela mesma razão, no caso das estratégias com unigramas ou bigramas, $|I| \leq 3$. No caso das estratégias com termos nominalizados e RLBs, $|I| \leq 18$, já que podem, em teoria, ser derivados até seis termos nominalizados (dois termos de cada palavra) e até 12 RLBs (combinação de seis termos, dois a dois, menos os 3 pares formados por termos diferentes mas derivados da mesma palavra). Substituindo em (27) é obtida, para estratégias com unigramas, bigramas e sintagmas nominais, a complexidade:

$$O(|C|+3+3+3^2) = O(|C|),$$

e para estratégias com termos nominalizados e RLBs:

$$O(|C|+3+3+18^2) = O(|C|).$$

Se for considerada a inclusão de operadores Booleanos, estes algoritmos apresentam a complexidade:

$$O(|C|)+O(|I|-1), \tag{28}$$

com $|I|-1$ operadores Booleanos incluídos.

Com o mesmo raciocínio anterior ($|I| \leq 18$ para estratégias com termos nominalizados e RLBs, e $|I| \leq 3$ para as outras estratégias), substituindo em (28) é obtida, para estratégias com unigramas, bigramas e sintagmas nominais, a complexidade:

$$O(|C|+2) = O(|C|),$$

e para estratégias com termos nominalizados e RLBs:

$$O(|C|+17) = O(|C|),$$

sendo $|C|$ a quantidade de caracteres do texto.

Logo, o algoritmo para *toquenização* da consulta apresenta complexidade da classe linear $O(n)$ e, em virtude das consultas curtas, os algoritmos para etiquetagem, normalização e preparação apresentam complexidade da classe constante $O(1)$. No total, então, o algoritmo para formulação da consulta, com ou sem inclusão de operadores Booleanos, apresenta complexidade da classe linear $O(n)$.

7.3.7 Pesquisa

A pesquisa de termos, no arquivo de índice, é realizada através do algoritmo da Figura 7.16. O mesmo algoritmo é utilizado para a pesquisa de bigramas.

Uma vez que, para consultas curtas como as utilizadas neste trabalho, $|T_C| \leq 3$, o algoritmo para pesquisa de termos apresenta a complexidade:

$$O(3|T|2) = O(|T|), \quad (29)$$

sendo |T| a quantidade de termos dos documentos.

No caso dos bigramas, a complexidade é:

$$O(|I|),$$

sendo $|I|$ a quantidade de descritores, ou seja, bigramas.

```
Para cada tC ∈ TC = {termos ou bigramas da consulta}
   Para cada t ∈ T = {descritores} /* indexados */
      Se tC = t então
         ler IDF /* se a estratégia exigir */
         ler lista de documentos e respectivas freqüências de t
         Interromper pesquisa em T
      Senão se tC < t (alfabeticamente) então
         Interromper pesquisa em T
```

Figura 7.16: Algoritmo para pesquisa de termos ou bigramas no arquivo de índice

Logo, o algoritmo para pesquisa de termos ou bigramas apresenta complexidade da classe linear $O(n)$.

```
Para cada sC ∈ SC = {sintagmas nominais da consulta}
   Para cada s ∈ S = {sintagmas nominais} /* indexados */
      Se sC está contido em s então
         ler IDF /* se a estratégia exigir */
         ler lista de documentos e freqüências
```

Figura 7.17: Algoritmo para pesquisa de sintagmas nominais no arquivo de índice

A pesquisa de sintagmas nominais no arquivo de índice é realizada através do algoritmo da Figura 7.17. O teste "`sC está contido em s`" realizado neste algoritmo, no pior caso, tem complexidade:

$$O(M_{Sc}(M_S - M_{Sc}+1)), \quad (30)$$

sendo M_{Sc} e M_S as quantidades médias de caracteres dos sintagmas nominais da consulta e do índice, respectivamente, sendo, no pior caso, $M_{Sc} = M_S/2$. Considerando (30), o algoritmo para pesquisa de sintagmas nominais apresenta a complexidade:

$$O(|S_C||S|(M_S/2)(M_S/2+1)). \quad (31)$$

Como $|S_C| \leq 3$, no caso das consultas curtas utilizadas, e como $|C| \geq |S|M_S$, $|C| \geq M_S/2+1$, para um texto com |C| caracteres, substituindo em (31) é obtida a complexidade:

$$O((3/2)|C||C|) = O(|C|^2),$$

sendo |C| a quantidade de caracteres do texto dos documentos.

Logo, o algoritmo para pesquisa de sintagmas nominais apresenta complexidade da classe quadrática $O(n^2)$.

A pesquisa de termos nominalizados e RLBs nos arquivos de índice é realizada através do algoritmo da Figura 7.18. A pesquisa dos termos nominalizados utiliza o algoritmo da Figura 7.16, com a substituição de "freqüência" por "evidência". No caso dos termos nominalizados, $|T_C| \leq 6$ já que, de cada palavra da consulta, podem ser derivados até 2 substantivos. Ainda assim, a complexidade, quanto aos termos, será aquela apresentada em (29).

```
Pesquisar termos nominalizados
Para cada r_C ∈ R_C = {RLBs da consulta}
   Para cada r ∈ R = {RLBs} /* indexadas */
      /* r_C = id_C(arg1_C,arg2_C) e r = id(arg1,arg2) */
      Se id_C = id e arg1_C = arg1 e arg2_C = arg2 então
         Ler lista de documentos e respectivas evidências de r
         Interromper pesquisa em R
      Senão
         Se (id_C = id e arg1_C = arg1 e arg2_C < arg2) ou (id_d = id e
         arg1_C < arg1) ou (id_C < id) (alfabeticamente) então
            Interromper pesquisa em R
```

Figura 7.18: Algoritmo para pesquisa de termos nominalizados e RLBs nos arquivos de índice

O algoritmo da Figura 7.18, então, apresenta a complexidade:

$$O(|T|)+O(|R_C||R|). \tag{32}$$

Considerando as consultas curtas utilizadas, $|R_C| \leq 12$, como já foi explicado. Como $|I| \geq |T|$ e $|I| \geq |R|$, já que $I = T \cup R$, substituindo em (32) é obtida a complexidade:

$$O(|I|+12|I|) = O(|I|),$$

sendo $|I|$ a quantidade de descritores (termos e relacionamentos).

Logo, o algoritmo para pesquisa de termos nominalizados e RLBs apresenta complexidade da classe linear $O(n)$.

7.3.8 Classificação

A classificação dos documentos recuperados é realizada através do algoritmo da Figura 7.19. Este algoritmo apresenta a complexidade:

$$O(|I_S||D_i|(|D_R|+1))+O(|D_R|)+O(|D_R|^2). \tag{33}$$

Como $|D_R| \geq |D_i|$ e $|I_S| \leq 18$ (nas estratégias com unigramas, o número de termos da consulta não ultrapassa 3 e, nas estratégias com termos nominalizados e RLBs, o número de termos e RLBs da consulta pode chegar, teoricamente, a 18), substituindo em (33) é obtida a complexidade:

$$O(18|D_R|^2+18|D_R|+|D_R|)+O(|D_R|^2) = O(18|D_R|^2+19|D_R|)+O(|D_R|^2) =$$
$$O(18|D_R|^2+19|D_R|)+O(|D_R|^2) = \tag{34}$$
$$O(|D_R|^2),$$

sendo $|D_R|$ a quantidade de documentos recuperados.

Logo, o algoritmo para classificação dos documentos recuperados apresenta complexidade da classe quadrática $O(n^2)$.

```
Para cada i ∈ I_S = {descritores selecionados na pesquisa}
   Ler IDF de i /* se a estratégia exige */
   Para cada d_i ∈ D_i = {documentos onde i ocorre}
      Para cada d_R ∈ D_R = {documentos recuperados}
         Se d_i = d_R então
            Atualizar dados para o cálculo do valor de relevância
            de d_R e interromper pesquisa em D_R
         Senão se d_i < d_R (em ordem de identificação) então
            Incluir d_i, antes de d_R, com dados para o cálculo do
            valor de relevância de d_i e interromper pesquisa em d_R
      Se d_i ∉ D_R então
         Incluir d_i ao final de D_R com dados para o cálculo do
         valor de relevância de d_i
Para cada d_R ∈ D_R
   Calcular valor de relevância de d_R
Ordenar D_R pelo valor de relevância
```

Figura 7.19: Algoritmo para classificação dos documentos recuperados

O algoritmo para classificação dos documentos recuperados com consulta Booleana, ao necessitar ordenar dois conjuntos de documentos recuperados (Figura 7.20), apresenta a complexidade adaptada de (34):

$$O(18|D_R|^2+37|D_R|)+O(|D_{R1}|^2+|D_{R2}|^2) = O(|D_R|^2+|D_{R1}|^2+|D_{R2}|^2). \quad (35)$$

```
Para cada i ∈ I_S = {descritores selecionados na pesquisa}
   Ler IDF de t /* se a estratégia exige */
   Para cada d_i ∈ D_i = {documentos onde i ocorre}
      Para cada d_R ∈ D_R = {documentos recuperados}
         Se d_i = d_R então
            Atualizar dados para o cálculo do valor de relevância
            de d_R e assinalar atendimento à consulta Booleana
         Senão se d_i < d_R (em ordem de identificação) então
            Incluir d_i, antes de d_R, com dados para o cálculo do
            valor de relevância de d_i e assinalar atendimento à
            consulta Booleana
      Se d_i ∉ D_R então inclui d_i ao final de D_R com dados para o
      cálculo do valor de relevância de d_i
Para cada d_R ∈ D_R
   Se d_R atende à consulta Booleana então
      Incluir d_R em D_R1 = {documentos recuperados no grupo
      superior}
   Senão incluir d_R em D_R2 = {documentos recuperados no grupo
   inferior}
   Calcular valor de relevância de d_R
Ordenar D_R1 pelo valor de relevância
Ordenar D_R2 pelo valor de relevância
```

Figura 7.20: Algoritmo para classificação dos documentos recuperados com consulta Booleana

Como $|D_R| = |D_{R1}|+|D_{R2}|$ e o pior caso acontece quando $|D_R| = |D_{R1}|$ e $|D_{R2}| = 0$ (ou $|D_R| = |D_{R2}|$ e $|D_{R1}| = 0$), substituindo em (35) é obtida a complexidade:

$$O(2|D_R|^2) = O(|D_R|^2), \quad (36)$$

sendo $|D_R|$ a quantidade de documentos recuperados.

Como $|D| \geq |D_R|$, substituindo em (36), é obtida a complexidade:

$$O(|D|^2),$$

sendo $|D|$ a quantidade de documentos da coleção.

Logo, o algoritmo para classificação dos documentos recuperados com consulta Booleana também apresenta complexidade da classe quadrática $O(n^2)$.

7.4 Tempo de processamento

Para a análise do tempo de processamento, feita a seguir, tanto em fase de indexação como em fase de busca, foram sempre utilizados os seguintes ambientes e recursos:

- computador Pentium III 866 MHz com 256Mb de RAM;
- sistema operacional Unix; e
- linguagem de programação C.

São analisados procedimentos da fase de indexação, quando os arquivos de índice são construídos, e da fase de busca, quando há a participação do usuário e as consultas são atendidas obtendo-se os documentos e classificando-os por relevância.

7.4.1 Em fase de indexação

São examinadas as mesmas estratégias de indexação descritas anteriormente na Seção 7.2.

No processo de indexação, porque foi executado sobre todo o conjunto de documentos em um único bloco de execução, não ocorreu a inclusão individual de cada documento em um arquivo de índice já construído. Tal inclusão se deu em estrutura de dados ainda em memória interna. A seguir é descrito esse processo de indexação da forma como foi realizado (em duas etapas) para a avaliação apresentada no Capítulo anterior:

- A *toquenização*, a etiquetagem e a normalização lexical constituíram o pré-processamento da coleção de documentos utilizada.
- Posteriormente, foi realizado o processo de indexação propriamente dito, específico de cada estratégia. Nesta segunda etapa, todas as estruturas de dados foram geradas em memória interna e, somente ao final do processamento de todos os documentos, os dados obtidos foram armazenados nos arquivos de índice.

Desta forma, não foram feitas distinções, quanto ao tempo de processamento, entre as diferentes etapas necessárias para a indexação. Para obter informações sobre o tempo gasto em cada procedimento da fase de indexação, estes procedimentos foram executados separadamente sobre 50 documentos aleatoriamente selecionados da coleção. Os dados apresentados na Tabela 7.2 são resultantes dessas execuções. Como os documentos utilizados possuem tamanhos diferentes, os resultados foram normalizados considerando a indexação de um documento com 1.000 itens de texto.

Obedecendo aos critérios de cada estratégia, os descritores dos 50 documentos indexados foram, então, incluídos, em cada caso, em arquivo de índice não vazio. Cada um desses arquivos, no momento da inclusão dos descritores do primeiro dos 50 documentos selecionados, contava com 4.106 documentos indexados, correspondentes à

coleção utilizada menos os 50 documentos a serem processados.

Os procedimentos de indexação apresentados na Tabela 7.2 são correspondentes aos algoritmos citados, para esta fase, na Seção 7.3. Não foram computados tempos de processamento para a etiquetagem de texto no caso de LM e ST porque, usualmente, os procedimentos de lematização e *stemming* não utilizam este recurso.

O tempo computado para o processo de nominalização (no caso de NM, BIGR e NMRL, ver Tabela 7.2) inclui o tempo de lematização. As palavras (adjetivos, substantivos e verbos, inclusive particípios) devem estar na forma de lemas para serem nominalizadas, conforme a estratégia adotada pela ferramenta CHAMA. O processo de nominalização em si, na verdade, é mais rápido que a lematização e o *stemmig*. Seu tempo é de 0,015s para processar um documento com 1.000 itens de texto.

Tabela 7.2: Tempo médio dos procedimentos de indexação de um documento com 1.000 itens de texto

estratégias	*toquenização*	etiquetagem	normalização	geração de descritores	inclusão no índice	total
VR	0,019s	–	–	0,010s	0,024s	0,053s
LM	0,019s	–	0,019s	0,009s	0,025s	0,072s
ST	0,019s	–	0,017s	0,008s	0,019s	0,062s
NM	0,019s	0,023s	0,034s	0,010s	0,019s	0,105s
SINN	0,019s	0,023s	0,019s	0,018s	0,028s	0,107s
BIGR	0,019s	0,023s	0,034s	0,017s	0,063s	0,155s
NMRL	0,019s	0,023s	0,034s	0,048s	0,454s	0,579s
média	0,019s	0,013s	0,022s	0,017s	0,090s	0,162s

O menor tempo de processamento cabe à estratégia VR e o maior à NMRL, de acordo com os dados da Tabela 7.2. Embora a nominalização e a identificação de RLBs contribuam para elevar o tempo de indexação na NMRL, o que mais contribui para esta elevação é o procedimento de inclusão dos descritores do documento corrente no arquivo de índice. Isto acontece porque as RLBs são estruturadas de forma a agilizar a pesquisa na fase de busca. Nessa estruturação, as RLBs são separadas pelo tipo (classificação, associação ou restrição) em arquivos diferentes e, em cada arquivo, são ordenadas, inicialmente, pelo identificador e, posteriormente, pelo primeiro argumento. Além de agilizar a pesquisa, essa estruturação diminui o espaço de memória necessário, porque identificadores iguais e primeiros argumentos iguais (com identificadores iguais) não são repetidos nos arquivos de índice.

7.4.2 Em fase de busca

Foram examinadas as principais estratégias de busca avaliadas no Capítulo anterior. Elas são descritas novamente a seguir.

VRf usa a estratégia de indexação VR, com termos não normalizados; LMf usa a estratégia de indexação LM, com termos lematizados; STf usa a estratégia de indexação ST, com termos na forma de *stems*; e NMf usa a estratégia de indexação NM, com termos nominalizados, todas com unigramas e baseadas em freqüência de ocorrência;

SINNf e SINNfB (esta com consulta Booleana) usam a estratégia de indexação SINN, com sintagmas nominais, BIGRf e BIGRfB (esta com consulta Booleana) usam a estratégia de indexação BIGR, com bigramas, btBIGRf usa a estratégia de indexação BIGR, com bitermos de acordo com a abordagem de Shrikanth e Srihari [SRI 2002], e mmNMRLf adota a abordagem de Changki Lee e Gary Lee [LEE 2005] e usa a

estratégia de indexação NMRL, com termos nominalizados (as RLBs são utilizadas apenas para identificar pares pai-filho de termos), todas baseadas em freqüência de ocorrência;

As estratégias sóRLBs, NMRLe e NMRLeB (esta com consulta Booleana), baseadas em evidência, usam a estratégia de indexação NMRL, com termos nominalizados (exceto sóRLBs, sem termos) e RLBs.

Na Tabela 7.3 são apresentados os tempos médios dos procedimentos para formulação de uma consulta com dois termos. Esses tempos podem ser atribuídos a cada estratégia de busca correspondente.

Tabela 7.3: Tempo médio dos procedimentos para formulação de uma consulta com dois termos

estratégias	*toquenização*	etiquetagem	normalização	preparação	total
VR	$0{,}04x10^{-3}$s	–	–	$0{,}02x10^{-3}$s	$0{,}06x10^{-3}$s
LM	$0{,}04x10^{-3}$s	–	$0{,}04x10^{-3}$s	$0{,}02x10^{-3}$s	$0{,}09x10^{-3}$s
ST	$0{,}04x10^{-3}$s	–	$0{,}03x10^{-3}$s	$0{,}02x10^{-3}$s	$0{,}09x10^{-3}$s
NM	$0{,}04x10^{-3}$s	$0{,}05x10^{-3}$s	$0{,}07x10^{-3}$s	$0{,}02x10^{-3}$s	$0{,}17x10^{-3}$s
BIGR	$0{,}04x10^{-3}$s	$0{,}05x10^{-3}$s	$0{,}07x10^{-3}$s	$0{,}03x10^{-3}$s	$0{,}19x10^{-3}$s
SINN	$0{,}04x10^{-3}$s	$0{,}05x10^{-3}$s	$0{,}04x10^{-3}$s	$0{,}04x10^{-3}$s	$0{,}16x10^{-3}$s
NMRL	$0{,}04x10^{-3}$s	$0{,}05x10^{-3}$s	$0{,}07x10^{-3}$s	$0{,}10x10^{-3}$s	$0{,}25x10^{-3}$s
média	$0{,}04x10^{-3}$s	$0{,}03x10^{-3}$s	$0{,}04x10^{-3}$s	$0{,}03x10^{-3}$s	$0{,}14x10^{-3}$s

As etapas de *toquenização*, etiquetagem e normalização são as mesmas etapas descritas para a fase de indexação. Na preparação são obtidos os termos e/ou relacionamentos e seus pesos e, conforme a estratégia, são incluídos os operadores Booleanos. O tempo para a inclusão de operadores Booleanos na consulta é menor que $0{,}01x10^{-3}$s, em qualquer estratégia.

Na Tabela 7.3 os dados se restringem à formulação da consulta. Na Tabela 7.4 são apresentados o tempo total da formulação e, ainda, os tempos de pesquisa e classificação em cada estratégia, considerando uma consulta com dois termos e a coleção de documentos utilizada neste trabalho. É possível observar que os tempos dos procedimentos para formulação de uma consulta com dois termos são ínfimos, quando comparados com os tempos gastos com os outros procedimentos da fase de busca.

Tabela 7.4: Tempo médio de processamento de uma consulta com dois termos

estratégias de busca	formulação	pesquisa	classificação	total
VRf	$0{,}06x10^{-3}$s	0,056s	0,083s	0,140s
LMf	$0{,}09x10^{-3}$s	0,046s	0,079s	0,125s
STf	$0{,}09x10^{-3}$s	0,042s	0,095s	0,137s
NMf	$0{,}17x10^{-3}$s	0,049s	0,081s	0,130s
SINNf	$0{,}16x10^{-3}$s	0,134s	0,076s	0,210s
SINNfB	$0{,}16x10^{-3}$s	0,134s	0,079s	0,213s
BIGRf	$0{,}19x10^{-3}$s	0,271s	0,086s	0,357s
BIGRfB	$0{,}19x10^{-3}$s	0,271s	0,088s	0,359s
btBIGRf	$0{,}19x10^{-3}$s	0,357s	0,080s	0,437s
mmNMRLf	$0{,}25x10^{-3}$s	0,131s	0,063s	0,194s
sóRLBe	$0{,}25x10^{-3}$s	0,091s	0,074s	0,165s
NMRLe	$0{,}25x10^{-3}$s	0,102s	0,076s	0,178s
NMRLeB	$0{,}25x10^{-3}$s	0,102s	0,078s	0,180s
média	$0{,}18x10^{-3}$s	0,137s	0,080s	0,217s

As estratégias com tempo de processamento de consulta acima da média são as que usam bigramas (ou bitermos) devido ao número de descritores que precisam pesquisar.

A estratégia mais rápida é LMf. Comparada com STf, ela gasta mais tempo em pesquisa, pois seu espaço de descritores é maior, mas é mais seletiva (recupera menos documentos), o que acarreta tempo de classificação menor. NMf gasta maior tempo na pesquisa do que LMf e STf, mas tem tempo total intermediário em relação às duas últimas, também devido a seu grau de seletividade.

Entre as estratégias que usam relacionamentos, a mais rápida é sóRLBe, e entre as que usam termos e relacionamentos, NMRLe é a que apresenta menor tempo. Esta estratégia é seguida de perto por NMRLeB, com 0,002s a mais, devido à consulta Booleana, tempo este gasto na classificação. A rapidez dessas estratégias de busca se deve à estruturação do espaço de descritores da estratégia de indexação NMRL.

7.5 Custo / Benefício

Levando em conta os custos decorrentes (i) do tamanho do espaço de descritores, ou seja, a memória necessária para armazenar os arquivos de índice construídos (memória externa), e (ii) dos tempos de processamento em fase de indexação e em fase de busca, pode-se ter uma idéia do benefício relativo de cada estratégia examinada.

Para representar a performance, quanto à recuperação dos documentos em atendimento às consultas, são utilizados os valores da medida *F* para os 10 primeiros documentos classificados.

Nas três Figuras seguintes, próximo ao posicionamento de cada estratégia é informada a respectiva relação custo/benefício. Também são desenhadas, em cada caso, as linhas de relação custo/beneficio média e com valor correspondente ao da estratégia NMRLeB, que implementa o modelo TR+ completo.

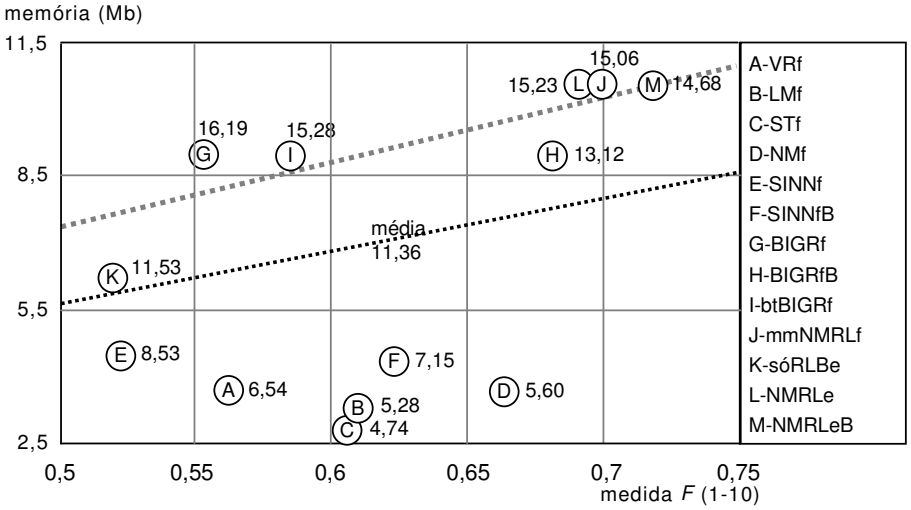

Figura 7.21: Espaço de memória e medida *F* (1-10)

Na Figura 7.21, o custo é medido pelo espaço de memória necessário para

armazenar os arquivos de índice originados a partir da coleção de documentos. As estratégias de busca com as quatro melhores relações quanto ao espaço de memória são as que usam unigramas: STf, LMf, NMf e VRf, nesta ordem. Próximas a este grupo estão as estratégias com sintagmas nominais. A linha de custo/benefício com valor 14,68 mostra a posição relativa de NMRLeB, na Figura 7.21. A relação média é 11,36. Ficam acima da média, também, a estratégia mmNMRLf e as que usam bigramas ou bitermos. A pior relação custo/benefício é da estratégia BIGRf.

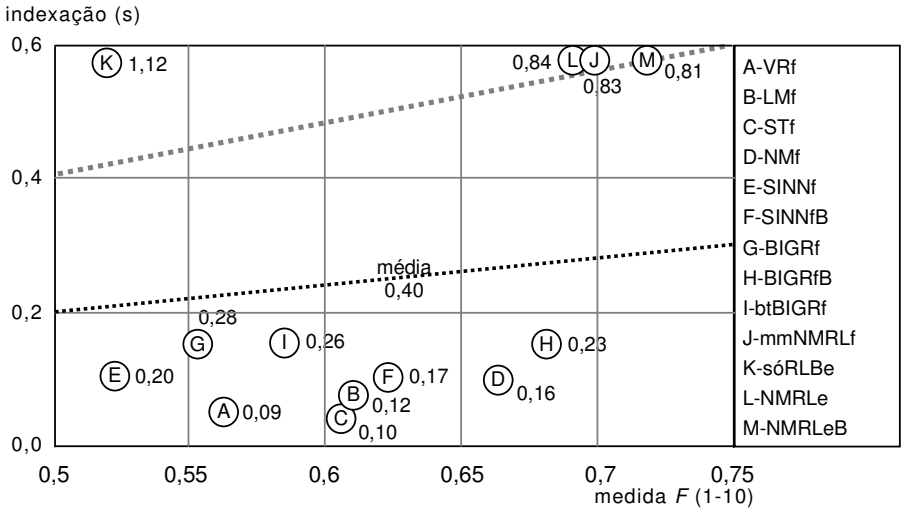

Figura 7.22: Indexação e medida *F* (1-10)

Na Figura 7.22, o custo é medido pelo tempo (em segundos) para indexação de um documento com 1.000 itens de texto. A linha de custo/benefício com valor 0,81 mostra a posição relativa de NMRLeB. A relação média é 0,40. Novamente, as quatro melhores relações são das estratégias que usam unigramas: VRf, STs, LMf e NMf, nesta ordem. Logo a seguir aparecem as estratégias com sintagmas nominais e bigramas ou bitermos. As estratégias que usam RLBs apresentam as piores relações custo/benefício.

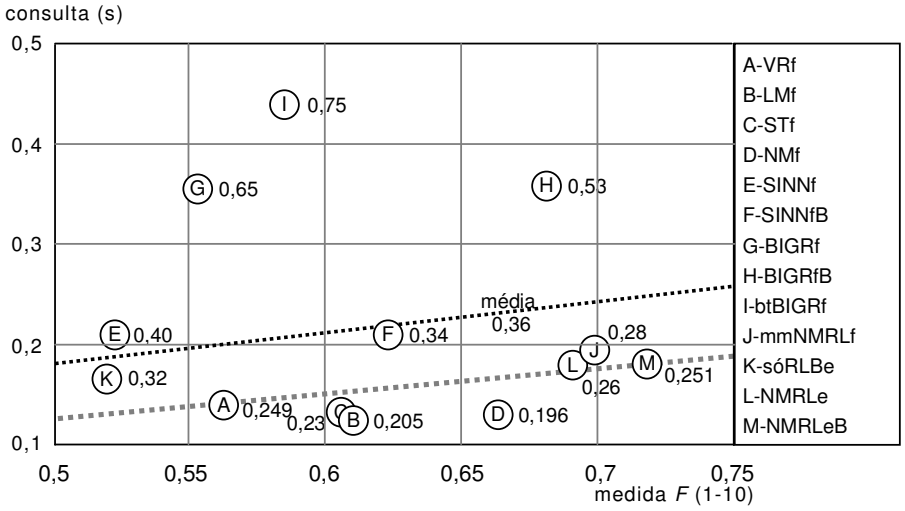

Figura 7.23: Processamento de consulta e medida *F* (1-10)

Na Figura 7.23, o custo é medido pelo tempo (em segundos) para processamento de uma consulta com dois termos. A linha de custo/benefício com valor 0,251 mostra a posição relativa de NMRLeB. A relação média é 0,36. Também aqui as estratégias com unigramas têm as quatro melhores relações: NMf, LMf, STf e VRf, nesta ordem. A quinta colocada é a estratégia NMRLeB. As estratégias que usam sintagmas nominais ficam próximas à média. As estratégias com bigramas ou bitermos apresentam as piores relações custo/benefício.

As estratégias que usam unigramas parecem ser imbatíveis quando a análise envolve custo/benefício, o que demonstra que há um preço a pagar pelo processamento da dependência de termos. Entretanto, é preciso levar em conta o que se perde em não considerá-lo. Perde-se em qualidade na classificação dos documentos relevantes em atendimento à consulta do usuário. Esta qualidade é obtida pela estratégia NMRLeB, representante do modelo TR+ proposto.

A relação custo/benefício da estratégia NMRLeB é elevada em fase de indexação, ficando acima da média. Por outro lado, em fase de busca, NMRLeB apresenta a menor relação custo/benefício entre todas as estratégias examinadas que consideram dependência de termos. Este fato é positivo porque a fase de busca está associada a uma característica crucial que não preocupa em fase de indexação: a presença do usuário.

Nas condições desta análise, os resultados obtidos demonstram que a relação custo/benefício da estratégia NMRLeB, que implementa o modelo TR+, é viável pelo menos para coleções de documentos como a de uma biblioteca digital com acervo de tamanho limitado. Sem a pretensão de determinar o limite de tal acervo, mas apenas como exercício de projeção, na Tabela 7.5 são apresentados os custos correspondentes à estratégia com melhor relação custo/benefício de cada aspecto analisado (memória, indexação e consulta) e à estratégia que implementa o modelo TR+ completo (NMRLeB) considerando uma coleção de documentos hipotética contendo edições equivalentes a 100 anos de um jornal de grande porte (como o Folha de São Paulo).

Tabela 7.5: Custos projetados para coleção de documentos hipotética correspondente a 100 anos de jornal de grande porte

estratégia	memória	indexação	consulta
de melhor custo/benefício	0,5Gb	3:30h	25s
do modelo TR+	2,0Gb	38:40h	35s

Na coluna “memória” da Tabela 7.5 é informado o espaço de memória necessário para armazenar os arquivos de índice (memória externa). A coluna “indexação” contém o tempo de processamento necessário para indexar pouco mais de 800.000 documentos. Na coluna “consulta” é informado o tempo de processamento para atender uma consulta com dois termos.

Estas são projeções lineares que levam em conta os dados e as condições dos experimentos realizados neste trabalho. Não são consideradas possíveis suavizações de curvas na evolução do crescimento do espaço de descritores (conforme Seção 7.2.1). Portanto, são considerados os piores casos para o espaço de memória necessário e para o tempo de consulta, já que essas projeções dependem do número de descritores. Para o tempo de indexação, por outro lado, como essa projeção depende da quantidade de palavras do texto (e não dos descritores gerados), os valores estimados devem estar mais

próximos da realidade.

7.6 Resumo do Capítulo

O custo de cada estratégia é considerado através (i) da memória necessária para armazenar os arquivos de índice, (ii) do tempo de processamento na fase de indexação, e (iii) do tempo gasto para atender a consulta, em fase de busca.

A estratégia mais econômica (em quantidade de descritores e, também, em tamanho do arquivo de índice) é a ST, que usa *stems*. As quantidades de termos em LM, com lemas, e NM, com termos nominalizados, são semelhantes.

Entre as estratégias que utilizam relacionamentos, a mais econômica é a SINN, com sintagmas nominais, tanto em espaço de memória quanto em número de descritores. As estratégias menos econômicas são a NMRL (com termos nominalizados e RLBs), em espaço de memória, e a BIGR (com bigramas), em número de descritores.

O crescimento do espaço de descritores, à medida que novos documentos vão sendo indexados, tende a ser suavizado em virtude de dois limites encontrados no léxico: (i) a quantidade de substantivos, pouco mais de 80.000 substantivos, e (ii) a quantidade de adjetivos, advérbios, substantivos, particípios e verbos, pouco mais de 140.000 palavras.

Tabela 7.6: Resumo da análise de complexidade – fase de indexação

estratégias	*tokeni-zação*	etique-tagem	norma-lização	geração de des-critores	inclusão no índice	Classe
VR	$O(\|C\|)$	–	–	$O(\|T\|^2)$	$O(\|T\|^2)$	$O(n^2)$
LM	$O(\|C\|)$	–	$O(\|P\|)$	$O(\|T\|^2)$	$O(\|T\|^2)$	$O(n^2)$
ST	$O(\|C\|)$	–	$O(\|P\|)$	$O(\|T\|^2)$	$O(\|T\|^2)$	$O(n^2)$
NM	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|T\|^2)$	$O(\|T\|^2)$	$O(n^2)$
SINN	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|P\|^2)$	$O(\|I\|^2)$	$O(n^2)$
BIGR	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|^2)$	$O(n^2)$
NMRL	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|^2)$	$O(n^2)$

Tabela 7.7: Resumo da análise de complexidade – fase de busca

estratégias de busca	tokeni-zação	etique-tagem	norma-lização	prepa-ração	pes-quisa	classi-ficação	Classe
VRf	$O(\|C\|)$	–	–	$O(\|T\|^2)$	$O(\|T\|)$	$O(\|D\|^2)$	$O(n^2)$
LMf	$O(\|C\|)$	–	$O(\|P\|)$	$O(\|T\|^2)$	$O(\|T\|)$	$O(\|D\|^2)$	$O(n^2)$
STf	$O(\|C\|)$	–	$O(\|P\|)$	$O(\|T\|^2)$	$O(\|T\|)$	$O(\|D\|^2)$	$O(n^2)$
NMf	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|T\|^2)$	$O(\|T\|)$	$O(\|D\|^2)$	$O(n^2)$
SINNf	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|P\|^2)$	$O(\|C\|^2)$	$O(\|D\|^2)$	$O(n^2)$
SINNfB	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|P\|^2)$	$O(\|C\|^2)$	$O(\|D\|^2)$	$O(n^2)$
BIGRf	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$
BIGRfB	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$
btBIGRf	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$
mmNMRLf	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$
sóRLBe	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$
NMRLe	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$
NMRLeB	$O(\|C\|)$	$O(\|P\|)$	$O(\|P\|)$	$O(\|I\|^2)$	$O(\|I\|)$	$O(\|D\|^2)$	$O(n^2)$

A análise de complexidade dos algoritmos das estratégias avaliadas conclui que

esses algoritmos apresentam classe de complexidade $O(n)$ ou $O(n^2)$. Um resumo desta análise para os algoritmos utilizados pelas estratégias avaliadas é apresentado para a fase de indexação, na Tabela 7.6, e para a fase de busca, na Tabela 7.7.

Na coluna "Classe", nessas Tabelas, é apresentada a classe de complexidade do conjunto de algoritmos em cada fase.

Para a análise de complexidade são considerados textos com |C| caracteres e |P| palavras, e espaços de descritores com |T| termos ou |*I*| descritores. Também é levado em conta que há na coleção |D| documentos.

Na fase de busca, quanto aos algoritmos para *toquenização*, etiquetagem, normalização e preparação, |C|, |P|, |T| e |*I*| referem-se ao texto da consulta e não ao texto dos documentos. A classe de complexidade desses algoritmos, em virtude do uso de consultas curtas, pode ser considerada $O(1)$, ou seja, constante.

Em fase de indexação, o menor tempo de processamento cabe à estratégia VR (sem normalização) e o maior à NMRL. O que mais contribui para a elevação do tempo de indexação na estratégia NMRL é o procedimento de inclusão dos descritores do documento corrente no arquivo de índice.

As estratégias com tempo de processamento de consulta acima da média são as que usam bigramas (ou bitermos) devido ao número de descritores que precisam pesquisar. A estratégia mais rápida é LMf, com termos lematizados. Das estratégias que usam termos e relacionamentos, as relacionadas ao modelo TR+ (NMRLe e NMRLeB) são as que apresentam menor tempo. Isto se deve à estruturação do espaço de descritores da estratégia de indexação NMRL.

As estratégias de busca com as quatro melhores relações quanto ao espaço de memória são as que usam unigramas, nesta ordem: STf, LMf, NMf e VRf. A relação de NMRLeB fica acima da média.

Quanto ao tempo de indexação, novamente as quatro melhores relações são das estratégias que usam unigramas, nesta ordem: VRf, STs, LMf e NMf. Ainda aqui, NMRLeB se mantém acima da média.

Para o processamento da consulta, as estratégias com unigramas também têm as quatro melhores relações, nesta ordem: NMf, LMf, STf e VRf. A quinta colocada é a estratégia NMRLeB, sendo a melhor entre as que consideram dependência de termos.

Há um preço a pagar pelo processamento da dependência de termos para se ter maior qualidade na classificação dos documentos relevantes. Esta qualidade é obtida pela estratégia NMRLeB, representante do modelo TR+ proposto. Com custo elevado em fase de indexação, esta estratégia, entretanto, tem relação custo/benefício baixa em fase de busca, quando o usuário está presente, o que a torna viável pelo menos para uma biblioteca digital com acervo de tamanho limitado, como uma coleção de documentos contendo as edições de 100 anos de um jornal de grande porte.

8 CONSIDERAÇÕES GERAIS

8.1 Introdução

Neste capítulo são discutidos aspectos importantes deste trabalho, constatados durante a pesquisa desenvolvida para a elaboração da proposta e no decorrer das experiências realizadas durante a avaliação do modelo TR+.

As principais diferenças entre a presente proposta e os modelos adotados pelos trabalhos correlatos pesquisados (apresentados na Seção 2.6) são destacadas em análise comparativa.

Causas de falhas na recuperação, pelas estratégias avaliadas, são levantadas e discutidas.

Diferenças entre as estratégias avaliadas são apontadas e discutidas através de exemplos ilustrativos das propriedades dos espaços de descritores. Essas propriedades estão relacionadas ao tamanho do espaço, à importância dos descritores, à geração de descritor único e à discriminação de conceitos distintos.

Finalmente, são sugeridas algumas características positivas para uma estratégia de RI. Essas características justificariam diferenças entre os resultados de recuperação das estratégias examinadas.

8.2 Diferenças entre a proposta e os trabalhos correlatos

Retomando a análise comparativa dos trabalhos correlatos iniciada na Seção 2.6.11, agora sob o enfoque das características mais importantes da proposta, são destacadas diferenças entre o modelo TR+ e os outros modelos com dependência de termos.

O uso de mais de um arquivo de índice, como ocorre no modelo TR+, não é comum entre os trabalhos pesquisados. Somente um dos trabalhos correlatos descritos (TCIM, do modelo TR) menciona esta estratégia. Outro (TCMM, do modelo TT) usa um arquivo de índice para os termos e, como os relacionamentos não são tratados como descritores, usa uma base de dados para identificá-los. No modelo TR+, arquivos de índice específicos, no formato de listas invertidas para os termos e para cada um dos tipos de RLBs, são usados para agilizar os procedimentos em fase de busca.

O modelo TR+ combina os modelos TR e RT no seguinte aspecto: tanto os pesos dos termos dependem de seus relacionamentos quanto os pesos dos relacionamentos dependem dos termos componentes. Tal dependência mútua se dá através do conceito de evidência. Nenhum dos trabalhos correlatos descritos apresenta esta característica. Quatro deles (TCIM, TCAB, TCTR e TCFM) seguem o modelo TR, onde as representatividades dos relacionamentos dependem dos termos; três deles (TCEI, TCPL e TCET) adotam o modelo RT, onde a representatividade dos termos é calculada a partir dos relacionamentos; dois trabalhos (TCNT e TCMM) seguem o modelo TT, onde a representatividade de um termo depende principalmente de sua freqüência de ocorrência, e um trabalho (TCBT) obedece ao modelo NG, onde a representatividade de um relacionamento depende da co-ocorrência dos termos componentes. Dos dez trabalhos pesquisados, apenas dois usam, como descritores, tanto

termos quanto relacionamentos, como o faz o modelo proposto aqui.

Todos os trabalhos correlatos apresentados utilizam estritamente métodos estatísticos para calcular o peso dos descritores. Os modelos NG, TT, TR e RT são baseados em freqüência de ocorrência. Ao adotar o conceito de evidência, é possível, para o modelo TR+, fazer com que termos e relacionamentos influenciem seus pesos mutuamente, pois esses pesos dependem da ocorrência de mecanismos de coesão frásica e não apenas de freqüência de ocorrência dos descritores.

Quanto à normalização lexical, entre os trabalhos pesquisados, a morfológica é mais comum que a léxico-semântica. A presente proposta realiza normalização léxico-semântica (assim como normalização morfológica) através da nominalização. No modelo TR+, a normalização léxico-semântica é realizada de forma restrita a alguns sinônimos de substantivos. Somente um dos trabalhos correlatos (TCNT, do modelo TT) implementa normalização léxico-semântica, e usa para isso um thesaurus de domínio. No modelo proposto não é utilizado um thesaurus, mas o conjunto de RLBs pode ser visto como uma estrutura que relaciona termos por classificação, restrição e associação (ver exemplos no Anexo D). O thesaurus simulado dessa forma, entretanto, não dispõe de relacionamentos correspondentes à sinonímia. Essa estrutura e seu uso são mencionados como trabalhos futuros na Seção 9.2.

A normalização sintática, em estratégias que usam dependência de termos, é feita usualmente através de regras gramaticais, no momento da geração dos relacionamentos, ou de regras transformacionais, exigindo esforço extra após a geração dos relacionamentos. Um dos trabalhos pesquisados (TCPL, do modelo RT) usa recursos estatísticos para definir a forma mais comum entre alternativas sintáticas possíveis. Assim como o modelo TR+ proposto, entre os dez trabalhos correlatos descritos, somente três (TCIM, TCTR e TCFM, todos do modelo TR) geram relacionamentos já no formato definitivo. Todos estes, assim como a proposta, o fazem através de regras gramaticais.

As RLBs não são rotuladas como ocorre, entre os trabalhos pesquisados, com os relacionamentos considerados por duas estratégias do modelo TR (TCAB e TCTR) e por outras duas do modelo RT (TCEI e TCET). Nessas estratégias, as relações recebem rótulos pré-definidos, como "*place*", "*time*", "*related-to*" e "*near-relation*".

Duas características mais marcantes do modelo TR+, que o diferenciam dos demais modelos, são o processo de nominalização e o conceito de evidência.

A nominalização distingue o modelo, ao ser usada como processo de normalização lexical que apresenta vantagens, como foi visto, sobre os usuais processos de lematização e *stemming*.

O conceito de evidência só é possível através da identificação de RLBs. Desta forma são obtidas informações úteis para o cálculo de peso dos descritores, além da usual freqüência de ocorrência.

8.3 Causas de falhas na recuperação

Os resultados de recuperação das estratégias examinadas são afetados por erros embutidos nos dados que elas recebem para processar ou por erros incluídos nos procedimentos executados por elas. Entre esses últimos, alguns erros são inerentes à abordagem escolhida e outros são decorrentes do processo desenvolvido para o procedimento executado.

Durante a experiência de avaliação realizada foram identificados os seguintes

tipos de erros (a notação colocada entre parênteses é usada nesta Seção):

- Etiquetagem incorreta (etq): uma palavra ou uma pontuação de um documento recebe uma etiqueta morfológica indevida.
- Falha da normalização (nor): diz respeito à incapacidade inerente ao tipo de normalização adotado e não a erro do processo aplicado. Esta falha pode (i) agrupar palavras não relacionadas ou (ii) não agrupar palavras da mesma família de significados. No caso de *stemming*, esses erros são denominados *overstemming* e *understemming*, respectivamente [FRA 92, ORE 2001].
- Erro do processo de lematização aplicado (lem).
- Erro do processo de *stemming* aplicado (stm).
- Erro do processo de nominalização aplicado (nom).
- Não tratamento de verbos (verb) nas estratégias com sintagmas nominais: impede a inclusão dessa categoria de palavras (normalizadas ou não) no espaço de descritores.
- Relacionamento indevido de termos (relac): palavras relacionadas na consulta não estão associadas em documentos tidos como relevantes, embora ocorram neles.
- Classificação indevida de documento por relevância (clas): é uma importante causa de erro na recuperação de documentos não relevantes. Ela ocorre porque são utilizados dois modos diferentes de classificação dos documentos: (i) o do julgamento de relevância, realizado por julgador humano, e (ii) o adotado pelas estratégias avaliadas. A classificação por julgamento humano é binária, ou seja, o julgador estabelece dois grupos de documentos: os relevantes e os não relevantes. Por outro lado, a classificação realizada pelas estratégias não é binária: os documentos são classificados em ordem decrescente do valor de relevância calculado. Se esse valor não é zero, o documento é recuperado e classificado. Tais documentos, recuperados como relevantes, possuem descritores que fazem com que o peso calculado não seja zero. Por outro lado, em alguns casos, o julgador humano pode considerar que, apesar da presença dos descritores, alguns documentos devem ser agrupados entre os não relevantes. Esses documentos são considerados, nesta análise, como classificados indevidamente. Nem sempre a classificação indevida deve ser vista como erro de recuperação, conforme é salientado adiante.
- Não tratamento semântico (sem): diz respeito ao não reconhecimento de relações semânticas. São incluídos, neste grupo, alguns casos de associações estatísticas de co-ocorrência não reconhecidas.

Na Tabela 8.1 são apresentadas as porcentagens de documentos recuperados não relevantes e as causas da recuperação dos mesmos pelas principais estratégias de busca examinadas. As porcentagens são calculadas levando em conta as quantidades de documentos recuperados apresentadas na coluna “rec”. No grupo de dados superior da Tabela 8.1 são considerados todos os documentos recuperados pelas estratégicas e, neste caso, a coluna “rec” informa a média dos documentos recuperados por consulta. No grupo inferior são analisados somente os 10 primeiros documentos recuperados por consulta em cada estratégia.

No caso da estratégia STf, sobressaem as quantidades de erros devidos a falhas inerentes da normalização (nor) e a erros do processo para o *stemming* aplicado (stm).

Um exemplo do primeiro caso é o das palavras "vestibular" e "vestíbulo", agrupadas, pela ferramenta utilizada, através do mesmo *stem* "vestibul". Um exemplo do segundo caso é o das palavras "rodovia" (substantivo) e "rodoviário" (adjetivo), que são derivadas erroneamente em *stems* diferentes: "rodov" e "rodovia", respectivamente.

Tabela 8.1: Causas da recuperação de documentos não relevantes

estratégias de busca	nor %	stm %	relac %	clas %	rec ñ rel %	rec
VRf	0,88	0,00	43,29	43,27	87,44	107,2
LMf	0,56	0,00	54,56	35,07	90,19	152,1
STf	9,92	6,02	24,18	52,82	92,94	233,0
NMf	0,51	0,00	41,64	50,03	92,18	207,0
SINNfB	0,37	0,00	26,89	62,05	89,31	138,9
BIGRfB	0,20	0,00	31,91	60,01	92,12	205,4
NMRLeB	0,23	0,00	32,74	59,20	92,17	207,1
média	1,81	0,86	36,46	51,78	90,91	178,7
VRf	0,00	0,00	15,80	7,40	23,20	10
LMf	0,00	0,00	11,00	8,00	19,00	10
STf	1,80	2,40	11,20	11,40	26,80	10
NMf	0,00	0,00	8,20	11,80	20,00	10
SINNfB	0,00	0,00	8,60	9,20	17,80	10
BIGRfB	0,00	0,00	6,20	12,40	18,60	10
NMRLeB	0,00	0,00	4,80	9,60	14,40	10
média	0,26	0,34	9,40	9,97	19,97	10

nor = falha inerente da normalização
stm = erro da estratégia de *stemming* adotada
relac = relacionamento indevido
clas = relevância indevida
rec ñ rel = documentos recuperados não relevantes
rec = documentos recuperados

No grupo superior da Tabela 8.1, salienta-se o fato de, em LMf, mais da metade dos documentos recuperados o são por relacionamentos indevidos (relac). No grupo de dados inferior da Tabela 8.1, é possível observar que as estratégias que consideram dependência de termos apresentam relacionamentos indevidos em menos de 10% dos documentos recuperados. O destaque fica com NMf, que também tem essa proporção inferior a 10%, já que, quanto às outras estratégias que não adotam unigramas, a redução dos relacionamentos indevidos deveria ser esperada. Neste sentido, NMRL é a mais eficiente. Relacionamento indevido ocorre, por exemplo, com a consulta "cinema brasileiro" quando o documento 1643 é pesquisado. Neste caso, a presença das palavras "brasileiros" e "cinema" (em "... conquistar os leitores brasileiros... Batman pode fazer sucesso nas telas de cinema ...") fazem com que algumas estratégias considerem indevidamente este documento como relevante para consulta exemplificada.

A classificação de relevância indevida (clas) pode ser exemplificada pela consulta "hotel". O julgamento humano de relevância realizado não considerou, por exemplo, o documento 1899 relevante para esta consulta. Entretanto, algumas estratégias recuperaram este documento para a consulta "hotel", devido ao trecho "O congresso, que se realizará no Hotel Hilton, espera reunir...". Classificações indevidas não devem ser consideradas indistintamente erros de recuperação no caso do grupo de dados superior da Tabela 8.1. Por outro lado, o mesmo não pode ser dito quanto ao grupo inferior, já que há limitação do número de documentos analisados a uma porção do topo da classificação: os 10 primeiros documentos. Entre estes deveriam estar os mais relevantes. O destaque, neste caso, fica para VRf que, apesar de não usar normalização

lexical nem relacionamentos, apresenta boa precisão no topo da lista dos documentos classificados por relevância, apesar dos 23,20% de documentos não relevantes entre os 10 primeiros.

Na Tabela 8.2, os dados dizem respeito à não recuperação de documentos relevantes.

Tabela 8.2: Causas da não recuperação de documentos relevantes

estratégias de busca	etq %	nor %	lem %	stm %	nom %	verb %	sem %	rel ñ rec %
VRf	0,00	16,42	0,00	0,00	0,00	0,00	3,65	20,07
LMf	0,46	9,63	0,28	0,00	0,00	0,00	0,46	11,40
STf	0,03	0,00	0,00	0,90	0,00	0,00	0,03	2,26
NMf	0,47	0,00	0,04	0,00	2,69	0,00	0,47	3,92
SINNfB	0,22	3,44	0,22	0,00	0,00	4,49	3,37	11,76
BIGRfB	0,02	0,00	0,02	0,00	1,91	0,00	1,96	3,92
NMRLeB	0,58	0,00	0,01	0,00	1,84	0,00	1,50	3,92
média	0,25	4,21	0,08	0,13	0,92	0,64	1,63	8,18

etq = erro de etiquetagem
nor = falha da normalização
lem = erro da estratégia de lematização
stm = erro da estratégia de *stemming*
nom = erro da estratégia de mominalização
verb = não consideração de verbos
sem = não tratamento semântico
rel ñ rec = documentos relevantes não recuperados

A estratégia NMRLeB é a mais prejudicada por erros de etiquetagem mas, mesmo assim, a porcentagem de documentos relevantes não recuperados, neste caso, não chega a 1%. Erros de etiquetagem podem ser exemplificados através da palavra "drible" no documento 1366, que deveria ter sido etiquetada como substantivo mas o foi como verbo. Essa etiqueta incorreta causou erro de lematização, pois foi gerado o lema "driblar" em vez do lema "drible". Também forçou a identificação de RLBs falsas, que se orientaram pela ocorrência do verbo, e impossibilitou a captura de sintagma nominal, já que, nesse caso, a palavra "drible" foi descartada por ser verbo. Outro exemplo, é o da palavra "animados" no documento 1402, que deveria ter sido etiquetada como particípio mas o foi como substantivo. A etiqueta incorreta, neste caso, causou erro de nominalização (o particípio não foi nominalizado) e, também, de identificação de RLB, pois o não reconhecimento do particípio alterou o seu relacionamento com o substantivo "desenhos", com o qual "animados" se vincula no texto.

A estratégia VRf é a que apresenta maior porcentagem de não recuperação de documentos relevantes por falha (no caso, ausência) de normalização. Ela é seguida pela estratégia LMf. Falhas de normalização, quanto a lematização, afetam diretamente as estratégias LMf e SINNf. Tais falhas não permite associar, por exemplo, palavras como "cirurgia" e "cirúrgico" ou "dançar" e "dança" ou "almoço" e "almoçar".

Também há erros específicos dos processos de normalização aplicados. Um exemplo de erro desse tipo na lematização (lem) é o que ocorre com o substantivo "deputada", cuja lematização não gerou a forma masculina ("deputado"), mantendo a forma feminina. Um exemplo de erro do processo de *stemming* aplicado (stm) é o relacionado com as palavras "franqueza" e "franquia", que apresentam o mesmo *stem* ("franq"). E um erro do processo de nominalização (nom) pode ser exemplificado através do particípio "animado", que deveria ter sido nominalizado como "animacao" e

não como "animo", conforme aconteceu.

O não tratamento de verbos (verb), específico da estratégia SINNf, pode ser exemplificado com o do trecho "o senador almoçou com o vice-presidente" do documento 898, onde os descritores considerados pela estratégia com sintagmas nominais são os termos "senador" e "vice-presidente". Assim, um termo correspondente ao verbo "almoçou" não pode ser incluído no espaço de descritores de SINNf.

Erros causados pela ausência de tratamento semântico (sem) são principalmente decorrentes da não utilização de recursos, como thesauri, do que propriamente por deficiência dos recursos adotados em cada estratégia. Um thesaurus semântico poderia identificar a associação que há entre palavras, como os substantivos "lider" e "liderança", por exemplo. A ausência ou a ineficácia da normalização léxico-semântica acarreta falhas na associação entre sinônimos ou quase-sinônimos, como "abuso sexual" e "violência sexual" ou "propaganda eleitoral gratuita" e "horário eleitoral gratuito". Outros casos, também incluídos aqui, são os que acontecem quando, por exemplo, a ocorrência do nome de um jogador de futebol em um documento não é suficiente para considerá-lo relevante a este esporte. Alguns casos como este último, onde a freqüência de co-ocorrência das palavras é importante, poderiam ser resolvidos através do uso de um thesaurus estatístico, por exemplo.

8.4 Propriedades dos espaços de descritores

Nesta Seção são discutidas características importantes dos espaços de descritores. Alguns conceitos apresentados aqui se encontram explícitos nos Capítulos anteriores, outros não. Todos são registrados nesta Seção de forma sistematizada para esclarecer as diferenças entre os distintos espaços de descritores examinados neste trabalho.

Dado um conjunto de conceitos *C* referentes a uma coleção de documentos, um espaço de descritores pode apresentar duas propriedades[20] principais:

- economia: a propriedade de possuir somente descritores necessários para descrever os conceitos de *C*, e
- representatividade: a propriedade de representar bem todos os conceitos de *C*.

Descritores desnecessários são os que descrevem conceitos ausentes ou já descritos por outros descritores. Para que um espaço de descritores seja mais econômico que outro, é preciso que tenha maior restrição na geração de descritores, com conseqüente redução da memória necessária, em fase de indexação, e do esforço de pesquisa a ser gasto, em fase de busca.

Entretanto, quanto maior a quantidade de conceitos descritos, maior será a cobertura do espaço de descritores. Neste sentido, os relacionamentos têm importante participação. A representação de dependências de termos através de relacionamentos tem o objetivo de aumentar a cobertura dos conceitos descritos. Os relacionamentos descrevem conceitos complexos que, de outra forma, não seriam descritos.

A representatividade de um espaço de descritores será maior, quanto maior for a sua cobertura e quanto mais preciso for o cálculo da representatividade de seus

[20] Essas duas propriedades são difíceis de medir, principalmente a representatividade que envolve o conceito de relevância. Alguns indicativos, entretanto, podem ser obtidos através de avaliações comparativas como as apresentadas no Capítulo 6 .

descritores em relação ao conteúdo do texto. Esta propriedade, portanto, pode ser entendida em nível de espaço de descritores e, também, em nível de cada descritor. As duas hipóteses seguintes podem fundamentar o raciocínio sobre a validade e a importância de um descritor.

Dado um documento d, os descritores i e j, e os respectivos pesos $W_{i,d}$ e $W_{j,d}$ em d,

(i) se $W_{i,d} > 0$, então i descreve um conceito presente em d, sendo i um descritor válido para d, e

(ii) se $W_{i,d} > W_{j,d}$, então o conceito descrito por i tem maior relevância que o conceito descrito por j em d, sendo i um descritor mais importante que j para d.

A validade e a importância de um descritor, portanto, dependem, respectivamente, da presença e da relevância do conceito descrito, e são medidas através de seu peso. O peso de um descritor é determinado diretamente pelo esquema de cálculo de peso utilizado mas, indiretamente, é afetado pela estratégia adotada para seleção e normalização do descritor.

A economia e a representatividade de um espaço de descritores dependem de alguns fatores (tamanho do espaço, representatividade dos descritores, geração de descritor único e discriminação de conceitos distintos) que são discutidos a seguir, a partir de exemplos. Esses exemplos, encontrados durante os experimentos realizados, ressaltam diferenças entre as estratégias avaliadas.

8.4.1 Tamanho do espaço de descritores

O tamanho de um espaço de descritores depende da capacidade de redução da quantidade de descritores e do tamanho de cada um. Tal redução pode resultar em economia de memória e em diminuição do esforço computacional na fase de busca.

A redução do tamanho do espaço de descritores é uma característica inerente ao processo de indexação. De acordo com os experimentos realizados, a eliminação de *stopwords* e de palavras duplicadas e a seleção/normalização dos descritores em cada estratégia causou uma redução de cerca de 99% de todo o conjunto de itens de texto da coleção de documentos utilizada (ver Tabela 8.3).

Tabela 8.3: Variações do tamanho dos espaços de descritores

	/ VR %	/ LM %	/ ST %	/ NM %	/ SINN %	/ BIGR %	/ NMR %	/ coleção %
VR	0,00	14,00	27,85	-1,08	-17,52	-58,91	-65,16	-99,70
LM	-12,28	0,00	12,15	-13,22	-27,64	-63,95	-69,44	-99,74
ST	-21,78	-10,83	0,00	-22,62	-35,48	-67,86	-72,75	-99,77
NM	1,09	15,24	29,24	0,00	-16,62	-58,46	-64,78	-99,70
SINN	21,24	38,21	54,99	19,93	0,00	-50,18	-57,76	-99,64
BIGR	143,34	177,41	211,10	140,72	100,72	0,00	-15,22	-99,28
NMRL	187,01	227,19	266,93	183,92	136,74	17,95	0,00	-99,15

Na Tabela 8.3 são apresentadas reduções (valores negativos) ou ampliações (valores positivos) dos espaços de descritores, comparados entre si e com a coleção de documentos, conforme cada uma das estratégia examinadas: VR, com variantes das palavras originais, LM, com lemas, ST, com *stems*, NM, com termos nominalizados,

SINN, com sintagmas nominais, BIGR, com bigramas, e NMRL, que segue o modelo TR+.

As ampliações das estratégias com relacionamentos em relação às que adotam unigramas ocorrem com o objetivo de aumentar a cobertura dos espaços de descritores, buscando descrever os conceitos correspondentes aos relacionamentos entre os termos.

A redução insuficiente pode resultar em falha na geração de descritor único, ou seja, na duplicação de descrição. Por outro lado, a redução excessiva pode resultar em ausência de descrição de algum conceito ou em descrição de mais de um conceito por um único descritor, ou seja, na perda da discriminação de conceitos distintos.

8.4.2 Representatividade dos descritores

A avaliação precisa da importância de cada descritor é crucial para a representatividade do espaço de descritores. A importância dada a um descritor em um documento destaca diferenças entre as estratégias avaliadas, quanto à seleção e/ou ao processo de normalização. Descritores que têm maior peso em uma estratégia podem, em outra, ter sua importância diluída ou podem não ser válidos em relação a um documento ou a toda a coleção.

A diluição da representatividade pode estar associada ao problema de geração de descritor único, discutido na próxima Seção. Se houver dois ou mais descritores de um mesmo conceito, a importância deles fica, em decorrência, diluída.

Por exemplo, considerando o documento 635, os pesos do termo "captação", em NM, e do *stem* "capt", em ST, são maiores que os pesos do termo "captação" em VR e LM. Como as palavras "captação", "captar" e "captou" ocorrem no documento 635, a importância da descrição do conceito associado a elas fica diluído em LM, pois há dois termos para representá-las ("captação" e "captar"), e mais ainda em VR, pois cada uma delas corresponde a um descritor. Em NM e em ST, porém, há apenas um termo representando as três palavras, não havendo diluição.

A variação do peso dos descritores pode ser ainda mais drástica, tornando-os não válidos para alguns documentos em algumas estratégias.

Por exemplo, o termo "acumulação" não é um descritor válido (tem peso zero), em VR e LM, para o documento 617. Por outro lado, este termo, em NM, e o *stem* "acumul", em ST, são válidos para o documento 617. Esses descritores foram derivados de quatro ocorrências do verbo "acumular" no referido documento.

O exemplo anterior se refere a um único documento, mas casos extremos podem acontecer quando um descritor, que aparece no espaço de descritores A, não é um descritor válido para documento algum no espaço B. Isto ocorre quando a estratégia que constrói B não deriva o referido descritor de nenhuma palavra da coleção de documentos.

Por exemplo, as palavras "abafamento" e "acirramento" (ou suas variantes) não ocorrem na coleção de documentos utilizada. Não são descritores válidos em VR e LM para qualquer documento. Contudo, em NM os termos "abafamento" e "acirramento" são descritores válidos para oito e dez documentos, respectivamente. Os *stems* dessas palavras têm comportamento similar em ST.

No caso dos relacionamentos, a análise das diferenças de representatividade entre as estratégias avaliadas é mais complexa porque eles são gerados em formatos diferentes (sintagma nominal, bigrama e RLBs). Entretanto, assim como acontece com os termos, a diluição da representatividade dos relacionamentos é decorrente da geração de descritor único, sendo discutida sobre este enfoque.

8.4.3 Geração de descritor único

Termos, que descrevem o mesmo conceito, podem ser incluídos em um espaço de descritores devido ao insucesso do processo de normalização lexical adotado, quando os termos apresentam diferenças morfológicas incomuns.

Por exemplo, para o documento 1021, o termo "risco", em VR e LM, e o *stem* "risc", em ST, não são descritores válidos porque o substantivo "risco" e suas variantes não ocorrem naquele documento. Contudo, para o mesmo documento, em NM, o termo "risco" é um descritor válido porque o processo de nominalização derivou este termo dos verbos "arrisquei" e "arrisco". Por outro lado, o *stem* "arrisc" – e não o *stem* "risc" – é um descritor válido, em ST, para o documento 1021. Neste caso, o conceito descrito tem um descritor único ("risco") em NM, e mais de um descritor ("arrisc" e "risc", este último com peso zero), em ST. Se forem executadas duas consultas, uma com o particípio "arriscado" e outra com o substantivo "risco", serão calculados dois valores de relevância diferentes para o documento 1021, com ST.

Outro exemplo é o caso do verbo "cobrir" e do substantivo "cobertura" no documento 4018. Em NM, o termo "cobertura" é um descritor único derivado de "cobrir" e de "cobertura". Por outro lado, no mesmo documento, os termos "cobrir" e "cobertura" existem com pesos distintos em VR e em LM. Em ST, os *stems* "cobr" e "cobert" existem também com pesos diferentes. Nesses três espaços de descritores (VR, LM e ST) não há um descritor único para o ato de cobrir algo.

Se, por um lado, um termo é descritor único de um conceito quando representa todas as variantes léxico-morfológicas associadas ao referido conceito, por outro, um relacionamento é descritor único quando representa todas as variantes sintáticas, principalmente quanto aos sintagmas nominais e às RLBs.

Os relacionamentos são descritores que têm maior especificidade quando comparados aos termos. Como cada relacionamento ocorre em alguns (poucos) documentos, o problema da geração de descrição única diminui, mas existe.

Por exemplo, a RLB "de(firmacao,acordo)" é um descritor único, em NMRL, para o conceito presente nos trechos: "acordo firmado", para os documentos 625 e 2981, "firmação de um acordo", para o documento 223, e "firmaram um acordo", para o documento 2720. Por outro lado são gerados os bigramas "(acordo,firmacao)", para os dois primeiros documentos, e "(firmacao,acordo)", para os outros dois, em BIGR. Em SINN são gerados os sintagmas "acordo firmado", para os documentos 625 e 2981, "firmacao de um acordo", para o documento 223, e "acordo", para o documento 2720.

Outro exemplo é o dos trechos "acusações feitas pelos sérvios", no documento 2854, e "sérvios acusam", no documento 2894. Para essas ocorrências, em NMRL, é gerada uma única RLB: "por(acusacao,servio)". Em BIGR é gerado o bigrama "(servio,acusacao)", para o documento 2894, e, quanto ao documento 2854, os termos "acusacao" e "servio" não são relacionados pois não são adjacentes. Em SINN, há um sintagma nominal constituído pelo trecho inteiro citado, quanto ao documento 2854, e apenas "servio" para o documento 2894.

8.4.4 Discriminação de conceitos distintos

A não discriminação de conceitos distintos causa ambigüidade na descrição. Nesses casos, dois ou mais conceitos são descritos indiscriminadamente por um só descritor.

Por exemplo, o mesmo *stem*, "caminh", é derivado dos substantivos "caminhada", para o documento 2427, e "caminhao", para os documentos 192 e 4134, em ST. Em NM,

entretanto, "caminhão" é um descritor válido para os dois últimos documentos, e "caminhada", para o primeiro. LM captura o termo "caminhao" para os documentos 192 e 4134, mas não considera "caminhada" um descritor válido para o documento 2427. Em VR, dos dois termos, somente "caminhao" é descritor válido e apenas para o documento 192.

Outro exemplo é a discriminação dos conceitos referentes a povo e publicação. Considere os seguintes trechos de dois documentos da coleção.

Documento 179: "... utilização de bens públicos ... pertence ao patrimônio público ..."

Documento 4093: "Seção publica 372 cartas este ano. Veículos publicou ..."

Para estes documentos, ST não discrimina povo de publicação, pois há um único *stem* ("public") derivado das palavras sublinhadas nos trechos apresentados. Por outro lado, em NM, é derivado o termo "povo" dos adjetivos do documento 179 e, a partir dos verbos do documento 4093, são derivados os termos "publicacao" e "publicador" (este último classifica a Seção "Veículos" do jornal que publicou as cartas). VR inclui todas as palavras sublinhadas como descritores, enquanto LM considera "publico", para o documento 179, e "publicar", para o documento 4093, como descritores válidos. Portanto, enquanto ST não discrimina os dois conceitos, NM, VR e LM discriminam (embora VR não o faça com descritores únicos).

São também bons exemplos os conceitos oriente (leste), nos documentos 2075, 2670 e 3050, e orientação, nos documentos 3001 e 3583. O primeiro pode ser representado pelo substantivo "oriente" e pelo substantivo (ou adjetivo) "oriental", e o segundo, pelo verbo "orientar" em algumas de suas conjugações, inclusive no particípio. LM não consegue associar o adjetivo "oriental" ao substantivo "oriente". VR falha com o termo "oriente", que pode estar associado a um conceito, sendo verbo, ou a outro, sendo substantivo. ST descreve os dois conceitos através do mesmo *stem*: "orient". NM consegue discriminar esses conceitos através dos descritores "oriente" e "orientação".

Outros casos podem, ainda, ser exemplificados com o adjetivo "livre", no documento 147, e o substantivo "livro", no documento 2382. Para estas palavras, ST gera um único stem: "livr". Do mesmo modo, o mesmo *stem*, "cobr", é derivado do substantivo "cobras", no documento 2281, e do substantivo "cobre" e do verbo "cobre", no documento 3756. Esses são alguns exemplos que explicam, em parte, a maior redução do espaço de descritores de ST, quando comparada com as demais estratégias.

No caso dos relacionamentos, há maior garantia de discriminação de conceitos distintos, pois os relacionamentos são inseridos no espaço de descritores, também, para diminuir a ambigüidade da descrição.

Por exemplo, as RLBs "de(democracia,alianca)" e "em(alianca,dedo)", os bigramas "(alianca,democracia)" e "(alianca,dedo)" e os sintagmas nominais "alianca democratica" e "alianca no dedo" discriminam os conceitos associados à "aliança" que ocorre nos documentos 2348 e 2347, representando um pacto democrático, e nos documentos 1589 e 3454, referindo-se a um anel no dedo.

Por outro lado, estratégias que adotam bigramas não conseguem distinguir "quarto com banheiro", do documento 1892, de "quarto sem banheiro", do documento 3841, ao gerar um único descritor "(quarto,banheiro)". Nestes casos, o descarte da preposição é crucial, impedindo a distinção dos conceitos envolvidos. A preposição está presente nas RLBs "com(quarto,banheiro)" e "sem(quarto,banheiro)" e nos sintagmas nominais "quarto com banheiro" e "diaria em quarto sem banheiro privativo", respectivamente, quanto aos documentos 1892 e 3841.

8.5 Características positivas para uma estratégia

Na Tabela 8.4 é apresentada a classificação geral, quanto aos resultados de recuperação, das principais estratégias de busca avaliadas. Para cada uma são informados a medida F para os 10 primeiros documentos recuperados, a precisão total, a revocação total e a quantidade média de documentos recuperados por consulta, levando em conta as 50 consultas processadas.

Tabela 8.4: Classificação das principais estratégias de busca

estratégias de busca	medida *F* 1-10	precisão 1-N	revocação 1-N	N
NMRLeB	0,718	0,078	0,961	207,1
mmNMRLf	0,700	0,079	0,961	206,1
NMRLe	0,692	0,078	0,961	207,1
BIGRfB	0,681	0,079	0,961	205,4
NMf	0,663	0,078	0,961	207,0
SINNfB	0,623	0,107	0,882	138,9
LMf	0,610	0,098	0,886	152,1
STf	0,606	0,071	0,977	233,0
btBIGRf	0,585	0,173	0,933	90,9
VRf	0,562	0,126	0,799	107,2
BIGRf	0,552	0,079	0,961	205,4
SINNf	0,522	0,107	0,882	138,9
sóRLBe	0,519	0,695	0,721	17,5
média	0,618	0,142	0,911	162,8

N = média de documentos recuperados por consulta

Da análise dos dados da Tabela 8.4 podem ser identificadas diferenças entre as estratégias, como causas de sucesso ou insucesso, e apontadas características positivas para obter bons resultados em RI.

É interessante observar que a estratégia STf, com *stems*, apresenta maior revocação total e menor precisão total. Em outro extremo está a estratégia sóRLBe, que possui menor revocação total e maior precisão total. Uma estratégia pode ser abrangente, quando recupera grande número de documentos, ou seletiva, quando classifica como relevantes poucos documentos. Quanto mais abrangente a estratégia (como STf), maior o risco de ser menos precisa. Ao contrário, quanto mais seletiva (como sóRLBe), maior o risco de ter revocação baixa.

O equilíbrio entre seletividade e abrangência é uma característica positiva? A estratégia sóRLBe é tipicamente seletiva porque utiliza apenas relacionamentos. Estes, em geral, repetem-se poucas vezes nos documentos, ou seja, são descritores mais raros e, em conseqüência, mais específicos que os termos. Outras estratégias são tipicamente abrangentes, como as que apresentam revocação superior a 0,9. Mais da metade delas possuem os maiores valores da medida *F* (ver Tabela 8.4). Isto parece indicar que, na prática, a abrangência é uma característica positiva, desde que acompanhada por outras características também positivas.

Das cinco estratégias abrangentes do topo da classificação apresentada na Tabela 8.4, as quatro superiores usam relacionamentos. Destas quatro, as três primeiras associam informações estatísticas e lingüísticas. Finalmente, destas três, a do topo da classificação combina essas informações com consulta Booleana.

De acordo com os resultados dos experimentos realizados aqui, são sugeridas, então, como características positivas de uma estratégia de RI, as seguintes: abrangência,

inclusão de relacionamentos no espaço de descritores, combinação de informações estatísticas e lingüísticas e uso de consulta Booleana.

8.6 Resumo do Capítulo

Algumas características diferenciam o modelo TR+ dos demais, considerando os trabalhos correlatos pesquisados. São elas: uso de mais de um arquivo de índice, cálculo do peso dos descritores baseado em evidência, uso de normalização morfológica e léxico-semântica através da nominalização, e uso de normalização sintática no momento da geração dos relacionamentos. Dentre estas características, as mais importantes são (i) a nominalização, que apresenta vantagens sobre a lematização e o *stemming*, e (ii) o conceito de evidência, que só é possível com a identificação de RLBs.

Os resultados de recuperação das estratégias examinadas são afetados por erros embutidos nos dados que elas recebem para processar ou por erros incluídos nos procedimentos executados por elas. Entre esses últimos, alguns erros são inerentes à abordagem escolhida e outros são decorrentes do processo desenvolvido para o procedimento executado.

No caso da estratégia STf, sobressaem as quantidades de erros devidos a falhas inerentes da normalização, e a erros do algoritmo adotado para *stemming*. Considerando todos os documentos recuperados, em LMf, mais da metade deles o são por relacionamentos indevidos. Considerando somente os 10 primeiros documentos recuperados, as estratégias que consideram dependência de termos apresentam relacionamentos indevidos em menos de 10% dos documentos, o que deveria ser esperado. Neste sentido, NMRL é a estratégia mais eficiente.

A estratégia NMRLeB é afetada por erros de etiquetagem mas o prejuízo é pequeno. A estratégia VRf é a que apresenta maior porcentagem de não recuperação de documentos relevantes por falha (no caso, ausência) de normalização. Ela é seguida pela estratégia LMf. Erros causados pela ausência de tratamento semântico (sem) são principalmente decorrentes da não utilização de recursos, como thesauri, do que propriamente por deficiência dos recursos adotados em cada estratégia.

A cobertura e a representatividade de um espaço de descritores depende de alguns fatores, como tamanho do espaço, importância dos descritores, geração de descritor único e discriminação de conceitos distintos.

A redução de um espaço de descritores resulta em economia de memória, em fase de indexação, e em menor esforço computacional, em fase de busca. A redução insuficiente pode prejudicar a geração de descritor único. Por outro lado, a redução excessiva pode resultar em perda da discriminação de conceitos distintos.

A diluição da representatividade pode estar associada ao problema de geração de descritor único. Se houver dois ou mais descritores de um mesmo conceito, a importância deles fica diluída.

Termos, que descrevem o mesmo conceito, podem ser incluídos em um espaço de descritores, devido ao insucesso do processo de normalização lexical adotado. Se, por um lado, um termo é descritor único de um conceito quando representa todas as variantes léxico-morfológicas associadas ao referido conceito, por outro, um relacionamento é descritor único quando representa todas as variantes sintáticas.

A não discriminação de conceitos distintos causa ambigüidade na descrição, fazendo com que dois ou mais conceitos sejam descritos indiscriminadamente por um só descritor. Há maior garantia de discriminação de conceitos distintos, no caso dos

relacionamentos, pois são inseridos no espaço de descritores também para diminuir a ambigüidade da descrição.

De acordo com os experimentos realizados aqui, são sugeridas, como características positivas de uma estratégia de RI, as seguintes: abrangência, uso de relacionamentos, combinação de informações estatísticas e lingüísticas e uso de consulta Booleana.

9 CONCLUSÃO

9.1 Principais contribuições

Quando um sistema de RI falha em discriminar conceitos ou em gerar descritores únicos ou, ainda, em medir a representatividade dos mesmos, essas falhas podem se transformar em sérios obstáculos para uma avaliação precisa do valor de relevância dos documentos.

Para tratar desses problemas, o modelo proposto com dependência de termos para RI combina abordagens estatísticas, lingüísticas e Booleanas. Considera termos e relacionamentos, na geração automática do espaço de descritores e na formulação Booleana da consulta, e adota cálculo de peso baseado em evidência, ao estabelecer a representatividade dos descritores. Neste sentido, o modelo TR+ apresenta, então, as seguintes características:

- uso de nominalização, como processo de normalização lexical na geração dos termos, para representar diferentes ocorrências léxico-morfológicas que são semanticamente equivalentes;
- identificação de RLBs dos tipos classificação, restrição e associação, que revelam relacionamentos entre termos, para representar diferentes ocorrências sintáticas que são semanticamente equivalentes;
- inclusão de termos nominalizados e RLBs no espaço de descritores, para obter maior cobertura de descrição de conceitos, incluindo relações semânticas importantes para a representação dos documentos e da consulta;
- uso de nova fórmula de cálculo do peso de termos e RLBs baseada em evidência para melhorar a representatividade do espaço de descritores, levando em conta, além da freqüência de ocorrência, a associação dos descritores a mecanismos de coesão frásica; e
- inclusão de operadores Booleanos na consulta, para complementar a especificação de dependências de termos.

Os experimentos de avaliação realizados procuraram demonstrar, através do modelo proposto e para a língua portuguesa, que:

- a nominalização, como processo de normalização lexical, pode propiciar melhores resultados de recuperação que os produzidos pelos processos tradicionais (lematização e *stemming*), conforme as informações apresentadas na Seção 6.2;
- a aquisição de informação lingüística através da identificação de RLBs pode contribuir positivamente para a descrição de dependências de termos, aumentando a cobertura do espaço de descritores, conforme as informações apresentadas nas Seções 6.4, 6.5 e 6.7;
- o cálculo da representatividade dos descritores baseado em evidência pode melhorar a classificação de relevância dos documentos, com vantagem sobre o cálculo baseado em freqüência de ocorrência, conforme as informações apresentadas na Seção 6.3;

- o uso de consultas com operadores Booleanos pode ser uma forma eficaz de complementar a especificação de dependências de termos, conforme as informações apresentadas nas Seções 6.5 e 6.7;
- a inclusão de conhecimento lingüístico, como a realizada no modelo proposto, pode apresentar relação custo/benefício viável na RI, conforme as informações apresentadas na Seção 7.5.

9.2 Trabalhos futuros

O modelo TR+ necessita de ferramentas para automatizar os processos de nominalização e de identificação de RLBs. Nos experimentos realizados aqui foram desenvolvidas e utilizadas as ferramentas CHAMA e RELLEX com esses objetivos. Ao contrário desses objetivos, as estratégias adotadas por essas ferramentas não caracterizam o modelo TR+. Assim, outros métodos para nominalização e identificação de RLBs podem ser pesquisados. Entretanto, em razão das estratégias adotadas terem contribuído positivamente para a validação do modelo proposto através da avaliação experimental realizada, são previstos entre os trabalhos futuros imediatos, sempre visando a otimização do tempo de processamento:

- inclusão de um módulo de etiquetagem de texto orientado às necessidades específicas do modelo TR+,
- adaptações na ferramenta de nominalização para torná-la independente da lematização e
- aperfeiçoamento das duas ferramentas utilizadas, visando aumentar precisão e abrangência de seus resultados.

Com esses recursos, o sistema estará pronto para ser testado em coleções de documentos mais extensas e com texto sem preparação prévia.

Outros trabalhos futuros de médio e longo prazos são descritos a seguir.

Além dos mecanismos de coesão frásica, outras ligações lingüísticas significativas [MIR 2003] merecem atenção quando o objetivo é qualificar o cálculo da evidência dos descritores e a identificação das RLBs. A coesão referencial, por exemplo, deve ser considerada através de resolução de anáforas, e a coesão lexical [STO 2004] deve ser incorporada com maior ênfase levando em conta relações semânticas, como sinonímia, hiponímia e meronímia.

Também merecem atenção os termos compostos. O reconhecimento desses termos pode diminuir o número de RLBs a serem identificadas. Nesses casos, há duas implicações:

- um conjunto de termos constituintes de uma composição deixa de produzir RLBs e passa a ser considerado como um termo único; e
- pode, assim, constituir um argumento em RLBs como qualquer termo.

O modelo deve ser testado em avaliação com consultas longas. Embora, os resultados da estratégia que implementa o modelo TR+ melhorem com consultas com dois ou três termos, quando comparados com aqueles produzidos por consultas com apenas um termo, isto é insuficiente para uma avaliação definitiva a respeito. Uma das razões que justificam tal teste é que as RLBs do tipo associação só podem ser identificadas em consultas longas. Constatou-se que as RLBs do tipo restrição contribuem efetivamente em sistemas de RI mas, quanto às outras relações, é necessário verificar se elas também podem contribuir neste sentido, ou se elas teriam papéis mais

adequados a desempenhar em outras aplicações. Por exemplo, as RLBs do tipo classificação podem ser testadas em sistemas de categorização de textos e as RLBs do tipo associação, em sistemas de sumarização automática.

O grafo formado pelo espaço de descritores que constitui o modelo TR+ pode ser entendido como um thesaurus de domínio, ou seja, extraído da própria coleção de documentos (ver exemplos no Anexo D). Tal thesaurus pode ser utilizado em:

- navegação para formulação de consulta, onde o usuário percorre os relacionamentos para selecionar os termos que deseja incorporar à sua consulta; e
- expansão automática de consulta, onde as RLBs são utilizadas para inserir termos relacionados àqueles originais da consulta formulada inicialmente pelo usuário.

O conceito de evidência, a partir de uma análise inicial, parece ser adequado a sistemas de sumarização automática de textos, com dois sentidos de pesquisa:

- as fórmulas para o cálculo da evidência de termos e RLBs podem ser utilizadas para, por exemplo, definir a evidência das sentenças de um texto. A evidência de uma sentença determinaria sua importância na construção do assunto tratado no texto, e determinaria sua seleção ou não para a geração do resumo;
- o grafo gerado pelas RLBs, conforme exemplos apresentados no Anexo D, pode ser usado para construir árvore(s) geradora(s) máxima(s), com arestas valoradas através dos pesos das RLBs. A(s) árvore(s) gerada(s) a partir de um texto pode(m) ser usada(s) para reconstruir as sentenças de modo seletivo, ou seja, descartando as arestas com menor evidência.

O conceito de evidência merece estudo para aplicação em sistemas de categorização de textos. O conceito de evidência pode ser utilizado para, de forma automática, identificar categorias importantes presentes na coleção de documentos. Essas categorias poderiam auxiliar na definição de conjuntos de documentos que evidenciam os mesmos assuntos ou assuntos equivalentes. Árvores geradoras máximas, mencionadas anteriormente, podem ser usadas nesta classificação: textos com árvores com um mínimo de diferenças pertenceriam à mesma categoria.

9.3 Publicações

São apresentadas, nesta Seção, as publicações relacionadas à presente tese.

A) RLBs e representação de texto

O artigo “Binary Lexical Relations for Text Representation in Information Retrieval” [GON 2005] apresenta aspectos do modelo proposto aqui, enfatizando as relações lexicais binárias e o conceito de evidência, e inclui uma avaliação comparativa do uso de termos nominalizados e RLBs com sintagmas nominais e com bitermos.

B) RLBs e hierarquia de temas

O artigo “Construção de hierarquia de temas e subtemas de texto” [GON 2003] apresenta uma contribuição para o reconhecimento de temas e subtemas de um texto. Para tanto, são construídas estruturas hierárquicas de descritores através de RLBs identificadas do texto.

C) Preposições em RLBs

O artigo “Mapping Syntactic Dependencies onto Semantic Relations”

[GAM 2002] apresenta o mapeamento das dependências sintáticas em relações semânticas. São analisados padrões gramaticais evidenciados através de preposições. Tais padrões são detectados através da localização das preposições em uma escala de papéis semânticos exercidos por complementos em RLBs.

D) Relações lexicais semânticas em expansão de consulta

Os artigos "Semantic Thesaurus for Automatic Expanded Query in Information Retrieval" [GON 2001], "Thesaurus com Estruturação Semântica e Operações Gerativas" [GON 2001a], e "Recuperação de Informação e Expansão Automática de Consulta com Thesaurus: uma avaliação" [GON 2001b] descrevem a construção de um thesaurus e seu uso em expansão automática de consulta para RI. São utilizadas, no processo de expansão de consulta, as relações lexicais semânticas definidas na estrutura Qualia da teoria do Léxico Gerativo. A avaliação positiva realizada foi o ponto de partida para a definição das RLBs como descritores de relações semânticas.

O artigo "Redefining Traditional Lexical Semantic Relations with Qualia Information" [GON 2004] define recursos aplicados na construção do thesaurus utilizado na experiência descrita no três artigos anteriores.

E) Substantivo como descritor

O artigo "Sintagma Nominal em Estrutura Hierárquica Temática na Recuperação de Informação" [GON 2001c] enfatiza a importância dos substantivos (núcleos dos sintagmas nominais) como descritores em RI.

F) Normalização lexical

O artigo "Normalização de itens lexicais baseada em sufixos" [GON 2003a] descreve uma estratégia para lematização baseada em sufixos e uma ferramenta que implementa esta abordagem.

9.4 Termos e relacionamentos em evidência

O modelo TR+ é um modelo de RI com utilização de PLN que apresenta procedimentos automatizados para as etapas da RI: geração de descritores, cálculo de pesos, formulação da consulta, pesquisa e classificação dos documentos recuperados. Os descritores são gerados a partir de procedimentos de nominalização e identificação de RLBs. Os pesos são calculados através do conceito de evidência. A consulta é formulada com a inclusão de operadores Booleanos predefinidos. A pesquisa é agilizada em arquivos de índice específicos para cada descritor, sendo as RLBs armazenadas de forma otimizada. Os documentos recuperados são classificados em dois grupos distintos de atendimento à consulta.

Todos esses procedimentos propostos foram motivados pela necessidade de se ter uma representação de texto mais rica que a usual concebida na indexação de documentos. Tal representação leva em conta a evidência dos descritores de acordo com a importância que apresentam no texto onde ocorrem na forma de palavras e frases. O modelo TR+ assume que a dependência dos termos estabelecida por mecanismos de coesão frásica dá pistas dessa importância, ou seja, revela os principais conceitos que o autor tem intenção de comunicar ao escrever seu texto.

REFERÊNCIAS

[ALL 2003] ALLAN, J.; KUMARAN, G. Stemming in the Language Modeling Framework. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 26., 2003. **Proceedings…** [S.l.: s.n.], 2003. p.455-456.

[ALL 95] ALLEN, J. **Natural Language Understanding**. Redwood City, CA: The Benjamin/Cummings Pub. Co., 1995. 654 p.

[ARA 97] ARAMPATZIS, A. T.; KOSTER, C. H. A.; TSORIS, T. IRENA: Information Retrieval Engine based on Natural Language Analysis. In: COMPUTER-ASSISTED INFORMATION SEARCHING ON INTERNET, RIAO, 1997. **Proceedings…** [S.l.: s.n.], 1997. p.159-175.

[ARA 2000] ARAMPATZIS, A. T. et al. Linguistically-motivated Information Retrieval. In: **Encyclopedia of Library and Information Science**. New York: M. Dehher, 2000. v. 69, p.201-222.

[ARA 2000a] ARAMPATZIS, A. T. et al. An Evaluation of Linguistically-motivated Indexing Schemes. In: BCS-IRSG – COLLOQUIUM ON IR RESEARCH, 2000. **Proceedings…** [S.l.: s.n.], 2000. p.91-111.

[BAE 99] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 513 p.

[BAR 2002] BARCALA, F. M. et al. Tokenization and Proper Noun Recognition for Information Retrieval. In: INTERNATIONAL WORKSHOP ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, DEXA, 13., 2002. **Proceedings…** [Los Alamitos]: IEEE Computer Society, 2002.

[BEA 91] BEARDON, C.; LUMSDEN, D.; HOLMES, G. **Natural Language and Computational Linguistics**. Inglaterra: Ellis Horwood, 1991. 232 p.

[BIC 2000] BICK, E. **The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. [S.l.]: Arthur University Press, 2000.

[BIS 2004] BISQUERRA, R.; SARRIERA, J. C.; MARTÍNEZ, F. **Introdução à Estatística – Enfoque informático com opacote estatístico SPSS.** Porto Alegre: Artmed, 2004. 255p.

[BRA 99] BRASCHLER, M.; RIPPLINGER, B. How Effective is Stemming and Decompounding for German Text Retrieval? **Information Retrieval Journal**, [S.l.], v. 7, p.291-316, 2004.

[BRO 2003] BRODER, A. et al. Efficient Query Evaluation using a Two-Level Retrieval Process. In: INT. CONF. ON INF. AND KNOWLEDGE MANAGEMENT, CIKM, 12., 2003. **Proceedings…** New York: ACM, 2003. p. 246-434.

[BRU 91] BRUZA, P. D.; WEIDE, Th. P. The Modelling and Retrieval of Documents using Index Expressions. **SIGIR Forum**, New York, v. 25, n. 2, p. 91-103, 1991.

[BUC 2004] BUCKLEY, C.; VOORHEES, E. M. Retrieval Evaluation with

Incomplete Information. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 27., 2004. **Proceedings…** [S.l.: s.n.], 2004. p.25-32.

[CAL 2004] CALLEGARI-JACQUES, S. M. **Bioestatística – Princípios e Aplicações**. Porto Alegre: Artmed, 2003. 255 p.

[CAM 2004] CAMPOS, J. **Os Enigmas do Nome**. Porto Alegre: Edipucrs, 2004.

[CEG 91] CEGALLA, D. P. **Gramática da Língua Portuguesa**. São Paulo: Nacional, 1991.

[CHO 68] CHOW, C.; LIU, C. Approximating discrete probability distributions with dependence trees. **IEEE Transations on Information Theory**, New York, v. 14, n. 3, p.462-467, 1968.

[CLE 67] CLEVERDON, C. W. The Cranfield tests on index language devices. In: ASLIB, 1967. **Proceedings…** [S.l.: s.n.], 1967. p. 173-192.

[COO 88] COOPER, W. S. Getting beyond Boole. **Information Processing and Management**, Oxford, v. 24, p.243-248, 1988.

[COO 94] COOPER, W. S. Some inconsistencies and misnomers in probabilistic information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 14., 1994. **Proceedings…** [S.l.: s.n.], 1994. p.57-61.

[CRO 91] CROFT, W. B.; TURTLE, H. R.; LEWIS, D. D. The use of phrases and structured queries in information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 14., 1991. **Proceedings…** [S.l.: s.n.], 1991. p.32-45.

[DIA 2000] DIAS, G. et al. Extraction Automatique d'Unités Lexicales Complexes: un Enjeu Fondamental pour la Recherche Documentaire. In: JACQUEMIN, C. (Ed.). **Traitement Automatique des Langues pour les Recherche d'Information**. Paris: Hermès Science Publications, 2000. p.447-493.

[DOB 98] DOBROV, B.; LOUKACHEVITCH, N. V.; YUDINA, T. N. Conceptual Indexing using Thematic Representation of Text. In: TEXT RETRIEVAL CONF., TREC, 6., 1998. **Proceedings…** [S.l.: s.n.], 1998. p. 403-454.

[FAG 87] FAGAN, J. L. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 10., 1987. **Proceedings…** [S.l.: s.n.], 1987. p.91-101.

[FÁV 97] FÁVERO, L. L. **Coesão e Coerência Textuais**. São Paulo: Ática, 1997.

[FER 99] FERREIRA, Aurélio B. de H. **Dicionário Aurélio Eletrônico** – Século XXI. Versão integral do Novo Dicionário da Língua Portuguesa – Século XXI. Rio de Janeiro: Nova Fronteira: Lexikon Informática, 1999.

[FRA 92] FRAKES, W. B.; BAEZA-YATES, R. **Information Retrieval:** Data Structures and Algorithms. New York: Prentice-Hall, 1992.

[GAM 2002] GAMALLO, P.; GONZALEZ, M.; AGUSTINI, A.; LOPES, G; LIMA, VERA L. S. de. Mapping Syntactic Dependencies onto Semantic Relations. In: WS ON NLP AND MACHINE LEARNING FOR ONTOLOGY ENGINEERING, ECAI, 2002. **Proceedings…** Lyon: [s.n.], 2002. p.15-22.

[GAO 2004] GAO, J.; NIE, J.; WU, G.; CAO, G. Dependence language model for information retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 27., 2004. **Proceedings…** [S.l.: s.n.], 2004. p.170-177.

[GON 2001] GONZALEZ, M.; LIMA, V. L. S. de. Semantic Thesaurus for Automatic Expanded Query in Information Retrieval. In: INTERNATIONAL SYMPOSIUM ON STRING PROCESSING AND INF. RETRIEVAL, 8., 2001, Santiago, Chile. **Proceedings…** Los Alamitos, CA: IEEE Computer Society Publications, 2001.

[GON 2001a] GONZALEZ, M.; LIMA, V. L. S. de. Thesaurus com Estruturação Semântica e Operações Gerativas. In: CONF. LATINOAMERICANA DE INFORMATICA, CLEI, 27., 2001, Mérida. **[Artículos]**. Mérida: Universidad de los Andes. 1 CD-ROM.

[GON 2001b] GONZALEZ, M.; LIMA, V. L. S. de. Recuperação de Informação e Expansão Automática de Consulta com Thesaurus: uma avaliação. In: CONF. LATINOAMERICANA DE INFORMATICA, CLEI, 27., 2001, Mérida. **[Artículos]**. Mérida: Universidad de los Andes. 1 CD-ROM.

[GON 2001c] GONZALEZ, M.; LIMA, V. L. S. de. Sintagma Nominal em Estrutura Hierárquica Temática na Recuperação de Informação. In: CONGRESSO DA SBC, 21.; ENCONTRO NACIONAL DE I. A., ENIA, 3., 2001, Fortaleza. **As tecnologias da Informação e a Questão social**: anais. Fortaleza: SBC, 2001. 1 CD-ROM.

[GON 2003] GONZALEZ, M.; MARTINS, A.; BARÃO, F.; LEITE, M.; LIMA, V. L. S. de. Construção de hierarquia de temas e subtemas de texto. In: BRAZILIAN SYMP. ON COMPUTER GRAPHICS AND IMAGE PROCESSING, SIBGRAPI, 16.; WS EM TECNOLOGIA DA INF. E LING. HUMANA, 1., 2003, São Carlos, SP. **Proceedings…** [S.l.: s.n.], 2003.

[GON 2003a] GONZALEZ, M.; TOSCANI, D.; ROSA, L.; DORNELES, R.; LIMA, V. L. S. de. Normalização de itens lexicais baseada em sufixos. In: BRAZILIAN SYMP. ON COMPUTER GRAPHICS AND IMAGE PROCESSING, SIBGRAPI, 16.; WS EM TECNOLOGIA DA INF. E LIN. HUMANA, 1., 2003, São Carlos, SP. **Proceedings…** [S.l.: s.n.], 2003.

[GON 2004] GONZALEZ, M.; LIMA, V. L. S. de. Redefining Traditional Lexical Semantic Relations with Qualia Information. **Revista Palavra**, [S.l.], n. 12, p.25-36, 2004.

[GON 2005] GONZALEZ, M.; LIMA, V. L. S. de; LIMA, J. V. de. Binary Lexical Relations for Text Representation in Information Retrieval. In: INT. CONF. ON APPLICATIONS OF NL TO INF. SYSTEMS, 10.; NLDB, 2005. **Proceedings…** [S.l.]: Springer-Verlag, 2005. p.21-31. (Lectures Notes in Computer Science, 3513).

[HAL 76] HALLIDAY, M.; HASAN, R. **Cohesion in English**. New York: Longman, 1976.

[HAR 95] HARMAN, D. K. The TREC Conferences. In: HYPERTEXT – INFORMATION RETRIEVAL – MULTIMEDIA, HIM, 1995. **Proceedings…** [S.l.: s.n.], 1995. p. 9-28.

[HIE 2000] HIEMSTRA, D. A probabilistic justification for using tfxidf term weighting in information retrieval. **International Journal of Digital Library**, [S.l.], v. 3, p.131-139, 2000.

[HIE 2002] HIEMSTRA, D. Term-Specific Smoothing for the Language Modeling Approach of Information Retrieval: The Importance of a Query Term. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 25., 2002. **Proceedings…** [S.l.: s.n.], 2002. p.35-41.

[HOU 2002] HOUAISS, A. **Dicionário Eletrônico Houaiss da Língua Portuguesa**: Versão 1.0.5. Rio de Janeiro: Objetiva, 2002.

[JAC 97] JACQUEMIN, C.; KLAVANS, J. L.; TZOUKERMANN, E. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In: AMACL, 35.; CONF. OF THE EUROPEAN CHAPTER OF THE ACL, 8., 1997. **Proceedings…** [S.l.: s.n.], 1997. p.24-31.

[JAC 99] JACQUEMIN, C.; TZOUKERMANN, E. NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax. In: STRZALKOWSKI, T. (Ed.). **Natural Language Information Retrieval**. [S.l.]: Kluwer Academic Publishers, 1999. p.25-74.

[JUR 2000] JURAFSKY, D.; MARTIN, J. **Speech and Language Processing** – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey, USA: Prentice-Hall, 2000. 934 p.

[KAT 2000] KATZ, B.; LIN, J. REXTOR: A System for Generating Relations from Natural Language. In: WS ON RECENT ADVANCES IN NLP AND IR, ACL, Hong-Kong, 2000. **Proceedings…** [S.l.: s.n.], 2000.

[KEH 2000] KEHDI, V. **Formação de Palavras em Português**. São Paulo: Ática, 2000.

[KOR 2004] KORENIUS, T. et al. Stemming and Lemmatization in the Clustering of Finnish Text Documents. In: CONF. ON INF. AND KNOWLEDGE MANAGEMENT, CIKM, 13., 2004. **Proceedings…** New York: ACM, 2003. p.625-634.

[KOW 97] KOWALSKI, G. **Information Retrieval Systems:** Theory and Implementation. Boston: Kluwer Academic Publishers, 1997. 282 p.

[KRO 93] KROVETZ, R. Viewing Morphology as an Inference Process. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 16., 1993. **Proceedings…** [S.l.: s.n.], 1993. p.191-202.

[KWO 98] KWOK, K. L.; CHAN, M. Improving Two-Stage Ad-Hoc Retrieval for Short Queries. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 21., 1998. **Proceedings…** [S.l.: s.n.], 1998. p. 250-256.

[LAN 69] LANCASTER, W. F. MEDLARS: Report on the evaluation of its operating efficiency. **American Documentation**, [S.l.], v. 20, p. 119-142, 1969.

[LAP 2002] LAPATA, M. The Disambiguation of Nominalizations. **Computational Linguistics**, [S.l.], v. 28, n. 3, p.357-388, 2002.

[LAH 2000] LAHTINEN, T. **Automatic Indexing**: An Approach using an Index Term Corpus and Combining Linguistic and Statistical Methods. 2000.

Tese (Doutorado) – Universidade de Helsinki, Finlândia.

[LEE 2005] LEE, C.; LEE, G. G. Probabilistic information retrieval model for a dependency structured indexing system. **Information Processing and Management**, Oxford, v. 41, p.161-175, 2005.

[LIN 2001] LIN, J. **Indexing and Retrieving Natural Language using Ternary Expressions**. 2001. Dissertação (Mestrado) – Massachusetts Institute of Technology, Cambridge, EUA.

[LIT 2000] LITKOWSKI, K. Question-Answering Using Semantic Relation Triples. In: TEXT RETRIEVAL CONF., TREC, 8., 2000. **Proceedings…** [S.l.: s.n.], 2000. p.349-356.

[LIU 2004] LIU, S. et al. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 27., 2004. **Proceedings…** [S.l.: s.n.], 2004. p. 266-272.

[LOO 97] LOSEE, R. M. Comparing Boolean and Probabilistic Information Retrieval Systems Across Queries and Disciplines. **Journal of the American Society for Information Science**, [S.l.], v. 48, n. 2, p.143-156, 1997.

[LOO 2001] LOSEE, R. M. Term Dependence: a basis for Luhn and Zipf Models. **Journal of the American Society for Information Science**, [S.l.], v. 52, n. 12, p.1019-1025, 2001.

[LOU 99] LOUKACHEVITCH, N. V.; SALLI, A. D.; DOBROV, B. V. Automatic Indexing Thesaurus Intended for Recognition of Lexical Cohesion in Texts. In: INT. CONF. ON APPLICATIONS OF NL TO INFORMATION SYSTEMS, 4., 1999. **Proceedings…** [S.l.: s.n.], 1999. p.203-208.

[LOU 2000] LOUKACHEVITCH, N. V.; DOBROV, B. V. Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. **Machine Translation Review**, [S.l.], n. 11, p.10-20, 2000.

[LYO 77] LYONS, J. **Semantics**. Cambridge: Cambridge University Press, 1977. 2 v.

[MAA 89] MAAREK, Y. S.; SMADJA, F. Full Text Indexing Based on Lexical Relations – An Application: Software Libraries. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 12., 1989. **Proceedings…** [S.l.: s.n.], 1989. p.198-206.

[MAC 73] MACIEL, J. **Elementos da teoria geral dos sistemas**. Petrópolis: Vozes, 1973.

[MAT 2000] MATSUMURA, A.; TAKASU, A.; ADACHI, J. The Effect of Information Retrieval Method Using Dependency Relationship Between Words. In: MULTIMEDIA INF. REPRESENTATION AND RETRIEVAL, RIAO, 2000. **Proceedings…** [S.l.: s.n.], 2000.

[MAR 2003] MARTINS, R.; NUNES, G.; HASEGAWA, R. Curupira: A Functional Parser for Brazilian Portuguese. In: PROPOR, 2003. **Proceedings…** [S.l.: s.n.], 2003. p.179-183.

[MEA 2000] MEADOW, C.T.; BOYCE, B.R.; KRAFT, D.H. **Text Information Retrieval Systems**. San Diego: Academic Press, 2000. 364 p.

[MIL 99] MIILLER, D. H., LEEK, T.; SCHWARTZ, R. 1999. A hidden Markov model information retrieval system. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 22., 1999. **Proceedings…** [S.l.: s.n.], 1999. p.214-221.

[MIR 2003] MIRA MATEUS, M. H. et al. **Gramática da Língua Portuguesa**. Lisboa: Ed. Caminho, 2003.

[MOE 2000] MOENS, M. F. **Automatic Indexing and Abstracting of Document Texts**. Boston: Kluwer Academic Publishers, 2000. 265 p.

[NAL 2002] NALLAPATI, R.; ALLAN, J. Capturing term dependencies using a language model based on sentence trees. In: INT. CONF. ON INF. AND KNOWLEDGE MANAGEMENT, CIKM, 11., 2002. **Proceedings…** [S.l.: s.n.], 2002. p.383-390.

[ORE 2001] ORENGO, V. M.; HUYCK, C. A Stemming Algorithm for the Portuguese Language. SYMP. ON STRING PROCESSING AND IR, SPIRE, 8., 2001, Chile. **Proceedings…** [S.l.: s.n.], 2001. p.186-193.

[PER 2000] PERINI, M. A. **Para uma Nova Gramática do Português**. São Paulo: Ática, 2000.

[PON 98] PONTE, J. M.; CROFT, W. B. A Language Modeling Approach to Information Retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 21., 1998. **Proceedings…** [S.l.: s.n.], 1998. p.275-281.

[PUS 95] PUSTEJOVSKY, J. **The Generative Lexicon**. Cambridge: The MIT Press, 1995. 298 p.

[RAW 92] RAWLINS, G. J. E. **Compared to What?** An Introduction to the Analysis of Algorithms. New York: Computer Science Press, 1992. 536 p.

[RAZ 2003] RAZERA, F. M.; ARAÚJO, M. P.; FRAGA, V. F. **Etiquetador Morfológico para o Português**. Trabalho de Conclusão. 2003. PUCRS, FACIN, Porto Alegre.

[RIJ 79] RIJSBERGEN, C.J. **Information Retrieval**. London: Bitterworths, 1979.

[ROB 76] ROBERTSON, S. E.; SPARCK-JONES, K. Relevance weighting of search terms. **Journal of the American Society for Information Sciences**, [S.l.], v. 27, n. 3, p.129-146, 1976.

[ROB 77] ROBERTSON, S. E. The Probability Ranking Principle in IR. **Journal of Documentation**, [S.l.], v. 33, n. 4, p.294-304, 1977.

[ROB 94] ROBERTSON, S. E.; WALKER, S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 17., 1994. **Proceedings…** [S.l.: s.n.], 1994. p.232-241.

[SAC 99] SACCONI, L. A. **Nossa Gramática:** Teoria e Prática. São Paulo: Atual, 1999. 576 p.

[SAL 68] SALTON, G.; LESK, M. E. Computer evaluation of indexing and text processing. **Journal of the Association for Computing Machinery**, New York, v. 15, p.8-36, 1968.

[SAL 75] SALTON, G.; WONG, A.; YANG, C. A Vector Space Model for Automatic Indexing. **Communications of the ACM**, New York, v. 18, p.613-620, 1975.

[SAL 82] SALTON, G.; BUCKLEY, C.; YU, C. T. An Evaluation of Term Dependence Models in Information Retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 5., 1982. **Proceedings…** [S.l.: s.n.], 1982. p.151-173.

[SAL 83] SALTON, G.; MACGILL, M. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

[SAL 86] SALTON, G. On the Use of Term Association in Automatic Information Retrieval. In: CONFERENCE ON COMPUTATIONAL LINGUISTICS, 11., 1986. **Proceedings…** [S.l.: s.n.], 1986. p.380-386.

[SAL 88] SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing and Management**, Oxford, v. 24, n. 5, p.513-523, 1988.

[SAV 2003] SAVARY, A.; JACQUEMIN, C. Reducing Information Variation in Text. **Text- and Speech-triggered Information Access**, [S.l.]: Springer, 2003. p.145-181. (Lectures Notes in Artificial Intelligence, 2705).

[SAV 96] SAVOY, J.; NDARUGENDAMWO, M; VRAJITORU, D. Report on the TREC-4 Experiment: Combining Probabilistic and Vector-Space Schemes. In: TEXT RETRIEVAL CONFERENCE, TREC, 4., 1996. **Proceedings…** [S.l.: s.n.], 1996. p. 537-548

[SHA 47] SHANNON, C. E. A Mathematical Theory of Communication. **The Bell System Technical Journal**, [S.l.], v. 27, 1948.

[SME 86] SMEATON, A. F. Incorporating Syntactic Information into a document retrieval strategy: an investigation. In: ANNUAL INT. ACM SIGIR CONF., 19., 1986. **Proceedings…** [S.l.: s.n.], 1986. p.103-113.

[SME 88] SMEATON, A. F.; RIJSBERGEN, C. J. van. Experiments on Incorporating Syntactic processing of User Queries into a Document Retrieval Strategy. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 11., 1988. **Proceedings…** [S.l.: s.n.], 1988. p.31-51.

[SON 99] SONG, F.; CROFT, B. A general language model for information retrieval. In: INT. CONF. ON INF. AND KNOWLEDGE MANAGEMENT, CIKM, 8., 1999. **Proceedings…** [S.l.: s.n.], 1999. p.316–321.

[SPA 97] SPARCK-JONES, K.; WILLET, P. (Ed.). **Readings in Information Retrieval**. California: Morgan Kaufmann Publishers, 1997.

[SPA 99] SPARCK-JONES, K. Information retrieval and artificial intelligence. **Artificial Intelligence**, [S.l.], v. 114, p.257-281, 1999.

[SPA 2000] SPARCK-JONES, K. ; WALKER, S. ; ROBERTSON, S. E. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments – Part 1 and 2. **Information Processing and**

Management, Oxford, v. 36, n. 6, p. 779-840, 1997.

[SRI 2002] SRIKANTH, M.; SRIHARI, R. 2002. Biterm language models for document retrieval. In: ANNUAL INTERNATIONAL ACM SIGIR CONF., 22., 2002. **Proceedings…** [S.l.: s.n.], 2002. p.425-426.

[STO 2004] STOKES, N. **Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain**. 2004. Tese (Doutoramento) – National University of Ireland, Dublin.

[STR 96] STRZALKOWSKI, T.; PEREZ-CARBALLO, J.; MARINESCU, M. Natural Language Information Retrieval in Digital Libraries. In: ACM INT. CONF. ON DIGITAL LIBRARY, 1., 1996. **Proceedings…** [S.l.: s.n.], 1996. p.117-125.

[STR 99] STRZALKOWSKI, T. et al. Evaluating Natural Language Processing Techniques in Information Retrieval. In: STRAZALKOWSKI, T. (Ed.). **Natural Language Information Retrieval**. [S.l.]: Kluwer Academic Publishers, 1999. p.113-145.

[SWA 88] SWANSON, D. R. Historical Note: Information Retrieval and the Future of an Illusion. **Journal of the American Society for Information Science**, [S.l.], v. 39, p.92-98, 1988.

[TOS 2002] TOSCANI, D. E.; ROSA, L. A.; DORNELES, R. C. S. **Normalizador de Itens Lexicais para o Português**. 2002. Trabalho de Conclusão. PUCRS, FACIN, Porto Alegre.

[VIE 2002] VIEIRA, R.; SALMON-ALT, S.; SCHANG, E. Multilingual Corpora Annotation for Processing Definite Descriptions. Advances in NLP, In: INTERNATIONAL CONF., PorTal, 3., 2002. **Proceedings…** [S.l.: s.n.], 2002. p.249-258.

[VIL 2002] VILARES, J., BARCALA, F. M.; ALONSO, M. A. Using Syntactic dependency-pairs conflation to improve retrieval performance in Spanish. In: **Computational Linguistics and Intelligent Text Processing**. [S.l.]: Springer-Verlag, 2002. p.381-390.

[VOO 2003] VOORHEES, E. M. Overview of TREC 2003.. TEXT RETRIEVAL CONFERENCE, 12., 2003. **Proceedings…** [S.l.:s.n.], 2003.

[WON 2000] WONDERGEM, B.; BOMMEL, P.; WEIDE, Th. P. Matching Index Expressions for Information Retrieval. **Information Retrieval**, [S.l.], v. 2, p.337-360, 2000.

[WON 2000a] WONDERGEM, B.; BOMMEL, P.; WEIDE, Th. P. Nesting and Defoliation of Index Expressions for Information Retrieval. **Knowledge and Information Systems**, [S.l.], v. 2, n. 1, 2000.

[ZHA 97] ZHAI, C. Fast statistical parsing of noun phrases of document indexing. CONFERENCE ON APPLIED NLP, 5., 1997. **Proceedings…** [S.l.: s.n.], 1997. p.312-319.

[ZIV 99] ZIVIANI, N.; BAEZA-YATES, R.; RIBEIRO-NETO, B. Text Operations. In: BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York : ACM Press, 1999. p.163-190.

ANEXO A AUTÔMATOS PARA NOMINALIZAÇÃO

São apresentados, neste Anexo, os autômatos finitos que implementam o processo de nominalização, para o Português, utilizados na ferramenta CHAMA.

Notação:

O formato das entradas dos autômatos é o seguinte:

s_η_1, η_2!

onde:

s é uma cadeia de caracteres a ser comparada com o trecho inicial da palavra a ser nominalizada, nos autômatos de exceções e de sinonímia, ou com trecho final, nos autômatos de padrões;

η_1 é uma operação de nominalização que deriva substantivos abstratos; e

η_2 é uma operação de nominalização que deriva substantivos concretos.

Se só for aceita a palavra completa coincidindo com a cadeia, então é usada a notação:

s_!η_1, η_2!

As operações de nominalização usam a seguinte notação (em formalismo EBNF):

[?ação1?]ação

onde:

? é um delimitador,

ação1, quando ocorre, é aplicável somente a adjetivos e

ação é aplicável a qualquer palavra ou, quando ação1 ocorre, somente a verbos.

As ações usam a seguinte notação (em formalismo EBNF):

–n+c | –n | +c | 0

onde:

n é o número de caracteres a ser eliminado do final da cadeia s;

c é o conjunto de caracteres a ser incluído no final da cadeia s; e

0 significa ausência de operação, ou seja, não há substantivo correspondente. Se for precedido de sinal de menos, significa “eliminar zero caracteres”.

As entradas dos autômatos são apresentadas, a seguir, em ordem alfabética e separadas por espaços para facilitar a consulta pelo leitor deste texto.

Para a execução dos autômatos, as entradas são armazenadas em uma estrutura de árvore ternária de pesquisa em cada autômato. Nessa estrutura cada nodo representa um estado do autômato, ou seja, um caractere da cadeia s.

A pesquisa de uma palavra na árvore ternária e a conseqüente nominalização determinada pelo autômato corrente obedece ao algoritmo apresentado na Figura A.1.

```
nodo atual = nodo Raiz da árvore ternária do autômato corrente
p = caractere inicial (ou final) da palavra de entrada
Enquanto há caractere atual p na palavra
      Se há nodo atual na árvore ternária
            n = caractere do nodo atual
            Se p<n alfabeticamente então
                  nodo atual = próximo nodo à esquerda
            Senão Se p>n alfabeticamente então
                  nodo atual = próximo nodo à direita
            Senão // p=n alfabeticamente
                  Se nodo atual corresponde a estado final então
                        Ler operações de nominalização
                        Incluir n na cadeia lida do autômato
                  nodo atual = próximo nodo ao centro
                  p = caractere posterior (ou anterior) da palavra
Se há operações de nominalização então
      Aplicar operações sobre a cadeia lida no autômato
Senão
      Não há nominalização possível neste autômato
```

Figura A.1: Algoritmo para nominalização em autômato finito

A seguir, são apresentadas as entradas dos autômatos para exceções (página 144), para padrões de adjetivos (página 155), para padrões de verbos (página 156) e para sinonímia (página 157).

Autômato para exceções

abacial_0,-4+de! abala_-1+o,+dor! abaluart_+amento,+ador! abana_-1+o,+dor! abandon_+o,+ador!
abdominal_0,-4+em! abelhud_+ice,+o! abenco_-6+bencao,+ador! abio_+se,0! abism_+amento,+ador!
abismal_0,-2+o! abissal_0,-3+mo! abjung_-1+cao,+idor! ablu_+cao,+ente! abnorm_-5+anormalidade,+al!
abob_+amento,+ador! abol_+icao,+idor! abord_+agem,+ador! abort_+o,+ivo! abraca_-1+o,+dor!
abrang_+encia,+edor! abrevia_+tura,+dor! abriga_-1+o,+dor! abrilhant_-9+brilho,+ador! abroch_+adura,+ador!
abrupto_0,0! abrut_+amento,+ador! absolut_-7+independencia,+o! absolv_+icao,+idor! abstem_+ia,+io!
abster_-1+ncao,-2+inente! absterg_-1+sao,+ente! abstergent_-3+encia,+e! absurd_+o,0! abund_+ancia,+ante!
abusa_-1+o,+dor! abusiv_-2+o,+o! acabar_-1+mento,-1+dor! acacal_+adura,+ador! academic_0,-1+a!
acalent_+o,+ador! acalma_-6+calma,+dor! acamp_+amento,+ador! acanave_+adura,+ador! acarea_+cao,+dor!
acarici_-7+caricia,+ador! acarinh_-7+carinho,-7+carinhoso! acarre_+io,+ador! acarret_+amento,+ador!
acasala_+mento,+dor! acautela_-8+cautela,+dor! aceita_+cao,+dor! aceitav_-1+cao,0! acena_-1+o,+dor!
acepilh_+adura,+ador! acera_+gem,+dor! acerb_+idade,0! acessa_-1+o,+dor! acetar_-2+ificacao,-2+ificador!
achar_!-1+do,-1+dor! acheg_+o,+ador! acidentar_-2+e,-1+dor! acinz_+entamento,+ador! acion_+amento,+ador!
acionari_0,-6+ao! aclama_+cao,+dor! acoberta_+mento,+dor! acoch_+o,+ador! acod_+amento,+ador!
acoim_+amento,+ador! aconchega_-1+o,+dor! aconselha_-9+conselho,+dor! acorrenta_+mento,+dor!
acostum_-7+costume,+ador! acre_!-1+idez,0! acredit_-7+crenca,+ador! acresce_-1+imo,+ntador! acritic_0,+o!
acrobatic_0,-3+cia! acromial_0,-2+o! actinomorf_+ia,0! acuar_-1+mento,-1+dor! acuda_+gem,+dor!
acudi_-4+juda,-4+judante! acula_+mento,+dor! acurar_-5+perfeicoamento,-1+dor! acurv_+amento,+ador!
adern_+amento,+ador! adetic_0,-3+s! adiante_!-7+dianteira,0! adiar_!-1+mento,-1+dor! adimpl_+emento,+idor!
adir_!-1+cao,-1+cionador! adivinh_+acao,+ador! adjuv_-4+juda,0! administrav_-1+cao,0! admira_+cao,+do!
admirav_-1+cao,0! admit_-1+ssao,-4+ceitador! admoesta_+cao,+dor! adocante_-3+mento,-3+dor!
adoece_-6+doenca,-6+doente! adoent_-6+doenca,0! adorav_-1+cao,0! adotar_-3+cao,-1+nte!
adquir_-5+quisicao,+ente! adua_+na,+neiro! adulterin_+idade,+o! adultero_-1+inidade,0! adunc_+idade,+o!
aduren_+cia,+te! advers_+idade,+o! adversario_0,0! advert_+encia,+idor! advir_!-5+origem,0!
advogar_-3+cacia,-1+do! aere_0,-3+r! aerobalistic_+a,+o! aeroespacia_0,0! aerostatic_-3+cao,+o! aetico_0,0!
afaga_-1+o,+dor! afamar_-6+fama,-1+dor! afana_-1+o,+dor! afe_+amento,+ador! afeic_+ao,+oador!
afer_+icao,+idor! aferr_+o,+ador! afet_?+o?+acao,+ador! afia_+cao,+dor! afianca_-7+fianca,+dor!
afim_!-1+nidade,-1+m! afinal_!0,0! afira_-3+ericao,-3+eridor! afiro_-3+ericao,-3+eridor! afli_+cao,+gidor!
aflora_+mento,+dor! aflu_+encia,+ente! afof_+amento,+ador! afofar_!-1+mento,-1+dor! afoga_+mento,+dor!
afoit_+eza,+o! afortuna_-8+fortuna,+dor! african_?0?+izacao,?-1?+izador! afrodi_0,0! afroux_+amento,+ador!
afugenta_+mento,+dor! afum_+adura,+ador! afunila_+mento,+dor! agente_-6+acao,0! agir_!-3+cao,-2+ente!
aglutina_+cao,+dor! aglutinav_-2+idade,0! agnatic_-3+cao,+o! agoni_+a,0! agonistic_+a,0! agracia_+cao,+dor!
agradar_-2+o,-1+dor! agrar_!-5+cultivo,-5+cultivador! agrav_+amento,+ador! agressiv_+idade,+o! agrest_+ia,+e!

agricol_-2+ultura,+a! agropecuari_0,+a! agrup_+amento,+ador! aguard_+o,+ador! agud_+eza,+o!
aguenta_7+resistencia,+dor! aguerri_+mento,+do! aidetic_-4+s,0! ajoelha_+cao,+dor! ajuda_-0,+nte!
ajust_+e,+ador! ajustav_-2+e,0! alacr_+idade,+e! alagadi_-2+mento,+co! alar_0,-4+asa!
alaranja_-8+laranja,+dor! alardea_-1,+dor! alarif_+agem,+ador! alarma_-1+e,-1+ista! alavanc_+agem,+ador!
albin_+ismo,+o! alcaguet_+agem,+a! alcanca_-1+e,+dor! alcool_!0,0! alcoolatr_-3+icidade,0! alcoolic_0,-2!
alcovit_+agem,+eiro! aleatori_+edade,+o! alegr_+ia,+ador! alergenic_0,-4+a! alerta_-0,+dor! alfabe_0,+to!
alfabetiz_+acao,+ador! alfandeg_?0?+agem,?+a?+ario! alfinet_+ada,+ador! algace_0,-2+s! algid_+ez,+o!
aliadofil_+ia,0! aliar_-1+nca,-1+do! alicerca_-1+e,+dor! aliena_+cao,+dor! aliment_?0?+acao,?+o?+ador!
alimp_+adura,+ador! alinhav_+o,+ador! alio_-1+anca,-1+ancador! alisa_+mento,+dor! almofad_+a,+ado!
alopra_+cao,dor! alouc_+amento,+ador! alpin_0,-2+es! altan_+aria,+eiro! alterco_-1+acao,-1+ador!
altern_+ancia,+ante! altiv_+ez,+o! alto_-1+ura,0! alucina_+cao,+dor! alucinog_0,-2+acao! alud_-1+sao,0!
aluga_-1+uel,+dor! alui_+cao,+dor! alumi_-5+iluminacao,-5+iluminador! alunar_-2+issagem,-2+issador!
alunissa_+gem,+dor! alvea_+mento,+dor! alvo_!-1+ura,0! amacia_+mento,+dor! amaldic_-7+maldicao,+oador!
amanhec_+er,+edor! amante_!0,0! amar_!-2+or,-1+nte! amarela_-1+o,-1+ecedor! amarelo_-1+idao,-1+ecedor!
amarg_+or,+ador! amaro_-1+gor,0! amarra_+cao,+dor! amassa_+mento,+dor! amazonic_0,-1+a!
ambicios_-3+ao,0! ambidestr_+ia,0! ambilev_+idade,+o! ameaca_-0,+dor! ameno_-1+idade,0! ami_+zade,+go!
amiasten_+ia,+ico! amigar_-3+zade,-2+o! amigav_0,-3+zade! amisto_0,-3+zade! amocamb_+amento,+ador!
amoj_+o,+ador! amolda_-6+moldagem,+dor! amorn_-5+mornanca,+ador! amostr_+a,+ador! amoux_+o,+ador!
ampar_+o,+ador! ampl_+itude,+o! amu_+amento,+ador! anabatic_0,-1+a! anacronic_-1+smo,+o! anal_!0,-2+us!
analfabet_+ismo,+o! analisa_-1+e,+dor! anarcossindical_+ismo,0! anarquis_+mo,+ta! ancestral_+idade,0!
anciao_!+nidade,0! ancor_+agem,+ador! andaluz_!0,+ia! andin_0,-2+es! androgin_+ia,0! anestesi_+a,+ador!
aneuploid_+ia,+e! anexo_-1+acao,0! angelical_!0,-7+jo! anglican_+ismo,+o! angulos_-1,0! angusti_+a,+ante!
anima_-1+o,+dor! animal_0,0! anisotrop_+ia,+ico! anisti_+a,+ador! aniversari_+o,+ante! anoitece_+r,0!
anojar_-6+nojo,-1+dor! anonim_+ato,+o! anorex_+ia,+ico! ansia_-1+edade,+so! ansio_-1+edade,+so!
antagonic_-1+smo,+o! antartico_0,0! antebraquial_0,-5+co! anti_0,0! antiamerican_+ismo,+o! antig_+uidade,0!
antimonar_+quismo,+quismo! antipod_+ismo,+a! antropomorf_+ia,+ico! anual_0,-3+o! anuir_-2+encia,-2+ente!
anular_!?0?-1+cao,-1+dor! apach_+ismo,+e! apaixo_-6+paixao,+nado! apalavr_+amento,+ador! apanh_+a,+ador!
apara_+gem,+dor! aparafus_+amento,+ador! aparenta_-2+cia,-1+e! apedreja_+mento,+dor! apegar_-2+o,-1+do!
apela_-1+o,+nte! apelid_+o,0! apendicular_0,-4+e! aperceb_-1+pcao,+edor! apetit_+e,0! apita_-1+o,+dor!
aplaca_+cao,+dor! aplaudi_-2+so,+dor! aplaus_+o,-s+didor! aplostav_-1,0! apocalip_0,+se! apocrif_+ia,0!
apodera_-7+posse,+dor! apofatic_-3+se,+o! apoiar_-2+o,-1+dor! apoline_0,-3+o! apolit_+ismo,+ico!
aporta_-1+e,+dor! aposent_+adoria,+ador! apossar_-7+posse,-1+dor! aposta_-0,+dor! aposto_!0,-2+icao!
apostolic_0,-3! apouc_+amento,+ador! aprazer_-7+prazer,0! apreciar_-1+cao,-1+dor!
apreen_?+sividade?+sao,+sor! apreensiv_-1+bilidade,0! aprend_+izagem,+iz! apressa_-7+pressa,+dor!
aprision_-8+prisao,+ador! apront_+o,+ador! apruma_-6+prumo,+do! apto_-1+idao,0! apua_+mento,+dor!
apunhala_-8+punhalada,+dor! apup_+ada,+ador! aquatic_0,-7+agua! arar_-4+cultivo,-4+cultivador!
arbitr_+agem,+o! arbitrari_+edade,+o! arbore_0,-4+vore! arbustivo_0,0! arcaic_+idade,+o!
arcar_-5+responsabilidade,-5+responsavel! arde_-1+or,+nte! ardilos_-2,0! areja_+mento,+dor! arenace_0,-4+ia!
arenos_-3+ia,0! arfa_+gem,+dor! argelin_0,-1+a! argentin_0,+a! argu_+icao,+idor! arido_0,0!
aristotel_+ismo,+ico! armadilha_-0,+dor! armar_-1+mento,-1+dor! arquiducal_0,-3+que! arquiepisco_0,-9+cebispo!
arquitet_+ura,+o! arranha_+o,+dor! arrecada_+cao,+dor! arreganh_+o,+ador! arregl_+o,+ador! arreliar_-1,-1+dor!
arremat_+e,+ador! arremed_+o,+ador! arremess_+o,+ador! arremet_+ida,+edor! arrepel_+ao,+ador!
arrepende_-1+imento,-1+ido! arrepi_+amento,+ador! arria_+mento,+dor! arrisca_-7+risco,+dor!
arrob_+amento,+ador! arroch_+adura,+ador! arroja_-1+o,+do! arrota_-1+o,+dor! arroxe_-6+roxo,+ador!
arruf_+o,+ador! arruin_-6+ruina,+ador! art_+e,+ista! arteria_0,0! artesanal_0,-1+to! artico_0,0! artistic_0,-5+e!
arvo_+amento,+ador! arvor_+agem,+ador! ascende_-2+sao,+nte! ascetic_0,0! asfalt_+amento,+ador!
asfixi_+a,_+ante! asiatic_0,-3! asila_-1+o,+dor! asperg_-1+sao,-1+sor! aspero_-1+eza,0! assa_+dura,+dor!
assalari_-8+salario,+ador! assalt_+o,+ante! assassin_+ato,+o! assedia_-1+o,+dor! assegura_-8+garantia,+dor!
assemelha_-9+semelhanca,+dor! assenti_+mento,+dor! assertivo_-4+cao,0! assest_+o,+ador!
assinala_+mento,+dor! assinar_-1+tura,-1+nte! assintomatic_0,0! assoberb_+amento,+ador! assobia_-1+o,+dor!
assola_+cao,+dor! assoma_+da,+dor! assombr_+o,+oso! assopr_+adela,+ador! assum_-1+ncao,+idor!
assust_-6+susto,+ador! astatic_-3+sia,+o! astigmatic_0,-1+a! astut_-1+cia,+o! atacant_0,-4+que!
atapet_+amento,+ador! atar_!-1+dura,-1+dor! ataref_+amento,+ador! atassalh_+adura,+ador!
atavi_+amento,+ador! atenc_+ao,+ioso! atende_-1+imento,+nte! atenta_+do,+dor! ater_!-1+ncao,-1+nto!
aterrador_!-9+terror,0! aterrar_-2+o,0! aterrissa_+gem,+dor! aterror_-7+terror,+izador! atesta_+do,+dor!
aticar_!-1+mento,-1+dor! atico_!0,0! atim_+ismo,0! atina_-5+raciocinio,+dor! atira_-5+tiro,+dor! ativo_-1+idade,0!
atlantico_0,0! atomic_0,+o! atorment_-8+tormento,+ador! atraca_+cao,+dor! atrapalha_+cao,+dor!
atrasa_-1+o,+dor! atrativ_+idade,+o! atrel_+agem,+ador! atreve_-1+imento,-1+ido! atrofia_-0,+dor!
atual_+idade,0! aturar_-6+resignacao,-1+dor1 auda_+cia,0! audio_0,0! aulic_+ismo,+o! aulic_0,-5+corte!
aument_+o,+ador! aureo_0,-5+ouro! auricular_0,-1! auroral_0,-1! ausenta_-2+cia,-1+e! auspicios_-1,0!
australia_0,0! austriac_0,-1! autentic_+idade,0! auto_0,0! autocritic_+a,+o! autograf_+o,+ador!
automobil_0,-3+vel! autonom_+ia,+o! autopromo_+cao,+tor! autopsia_-0,+dor! autor_!+ia,0! autoral_0,-2!
autoritar_+ismo,+io! autoriza_+cao,+dor! auxiliar_-2+o,0! avanca_-1+o,+dor! aventur_+a,+eiro!
averb_+amento,+ador! averigua_+cao,+dor! avess_+ia,0! aviar_-1+mento,-1+dor! avico_0,-3+e!
avilta_+mento,+nte! avista_6+vista,+dor! avizinh_-7+vizinhanca,-7+vizinho! avoeng_0,-3! axial_0,-5+eixo!
axilar_0,-1! azar_?0?+acao,?-0?+ador! azedo_-1+ume,0! azul_0,0! azula_+mento,+dor! babar_-1,-1+dor!
bagunc_+a,+eiro! baian_0,-1! bailar_-2+e,+ino! baita_-5+grandeza,0! baixa_-5+abaixamento,+dor!
baixo_-1+eza,0! bajar_-5+vagem,-1+dor! balanca_-1+o,+dor! balistic_+a,0! balof_+ice,+o! balsamic_0,-2+o!
bancar_-6+assuncao,-1+dor! bancari_0,-3+o! banhar_-2+o,-2+ista! barato_-1+eza,0! barbar_-1,-1+do!
barbaro_-1+ismo,0! barbea_+da,+dor! barbeir_+agem,+o! barbitur_+ismo,+ico! barganha_-0,+dor!
barrig_+a,+udo! barroco_-2+quismo,0! barros_0,-1! barulhent_-3+o,+o! basbac_-1+quice,+a! basea_-1,+do!
basta_-5+suficiencia,+nte! bastard_+ia,0! basto_-1+idao,0! batalh_+a,+ador! bate_-1+ida,+dor!
batiza_-2+smo,+ador! batuc_+ada,+ador! beb_+edeira,+ado! bebado_-3+edeira,0! bebedo_-1+eira,-3+ado!
beber_-2+ida,-1+dor! beberrao_-4+deira,0! beijar_-2+o,-1+dor! beira_-0,+do! belic_0,-5+guerra!
beliger_+ancia,+ante! belisca_+o,+dor! belo_!-1+eza,0! bendit_-3+cao,+o! benefic_+io,+iador!
benevol_+encia,+o! berra_-1+o,+nte! besunt_+adela,+ador! biblic_0,-1+a! bicicle_0,+ta! bicud_+ez,+o!

binocular_0,0! bioestatistic_+a,0! biograf_+ia,+o! bionic_0,+o! bipartir_-1+cao,-1+dor! bisar_-2,-1+dor!
bisbilhot_+ice,+eiro! bisonh_+ice,+eiro! bizarr_+ia,0! blef_+e,+ador! blind_+agem,+ador! bloca_+gem,+dor!
bloque_+io,+ador! bobea_+da,-2+o! bobin_+agem,+ador! boemi_+a,+o! boicot_+e,+ador! bolar_-1+cao,-1+dor!
bolivian_0,-1! bolor_+encia,+ento! bom_-1+ndade,-1+m! bombarde_+io,+ador! bombastic_0,0!
bombe_+amento,+ador! bonach_+eirice,+ao! bondos_-2+ade,+o! bonit_-4+eleza,+o! boquea_+da,+dor!
borbulh_+a,+ante! borda_+do,+dor! borrach_-7+bebedeira,+o! borrar_-1+da,-1+dor! botanic_0,+a!
botar_-5+colocacao,-5+colocador! bovin_0,-3+i! bracal_0,-2+o! brace_+agem,+ador! branc_+ura,-1+queador!
brandi_+mento,+dor! brando_-1+ura,0! branquial_0,-1+s! braquidactil_+ia,0! brasileir_0,-3! brav_+ura,+o!
brea_+dura,+dor! breca_+da,+dor! brechar_-1,-1+dor! brev_+idade,+e! breveta_-2+,+do! briga_-0,+dor!
briguent_-4+a,+o! brilh_+o,+ante! brinca_+deira,+lhao! brinda_-1+e,+dor! brio_-0,+so! briquet_+agem,+eiro!
britanic_0,-8+gra-bretanha! broch_+ura,+ador! bronc_-1+quice,+o! brosl_+adura,+ador! brumal_0,-1!
brusc_-1+quidao,+o! bruto_-1+alidade,-1+o! brux_+ismo,+o! bruxul_+eio,+ador! bucal_0,-4+oca!
bucol_+ismo,+ico! bufar_!-2+o,-1+dor! bufo_!+naria,0! bugr_+ismo,+e! bulatic_0,-3! burgues_+ia,-1+s!
burlar_-1,-1+dor! burlesc_-7+comicidade,+o! burro_-1+ice,0! busca_-0,+dor! buzina_+cao,+dor!
cabece_+io,+ador! cabelu_-1+o,0! cabiv_-1+mento,0! cabocl_+ismo,+o! cabotin_+ismo,+o!
cabrar_-1+gem,-1+dor! cabroc_+a,+ador! cacar_-1+da,-1+dor! caceta_+da,+dor! cachop_+ice,+eiro!
cachorr_+ismo,+o! cadencia_-1+a,+dor! caduc_+idade,+o! cafon_+ice,+a! cagar_-1+da,-1+dor!
caiar_-1+cao,-1+dor! cainh_+eza,+o! cair_!-4+queda,-1+dor! caititu_+agem,+ador! cala_+dura,+dor!
calaceir_+ice,+o! calamit_-1+dade,0! calandr_+agem,+ador! calcaneo_0,-2+har! calcul_+o,+ador!
calendar_0,-1+s! calhar_!-6+conveniencia,-1+dor! calm_+a,+ante! calore_-1,+nto! calour_+ice,+o!
caluni_+a,+ador! camb_+adela,+ador! cambalea_+da,+nte! cambia_?0?-1+o,?-1+o?+dor! caminha_+da,+nte!
campal_0,-2+o! campestre_0,-5+o! camufl_+agem,+ador! canalar_0,-2! canarinh_0,+o! canaviei_0,-4!
cancerig_0,-2! candidat_+ura,+o! candido_-3+ura,0! canela_+gem,+dor! canfor_+ismo,+ico! canguinh_+ez,+o!
canibal_+ismo,0! canicular_0,-1! cansa_+co,+do! canta_-2+cao,-1+or! caotic_-3+s,0! capach_+ismo,+o!
capaz_-1+cidade,0! capilicial_0,-2+o! capin_+a,+ador! capital_-7+importancia,0! capota_+gem,+dor!
caprich_+o,+oso! caprichos_-1,0! capsular_0,-1! caracteristic_+a,0! caracterizav_-1+cao,0! carbonic_0,-2+o!
carcer_+agem,+eiro! carcerari_0,-3+e! carche_+io,+ador! carcom_+icao,+idor! cardeal_0,0!
cardiaco_0,-7+oracao! cardial_0,-1! cardiovascular_0,0! carece_-2+ncia,-2+nte! caret_+ice,+a!
cariar_-2+e,-1+dor! caricat_+ura,+urador! carimb_+agem,+ador! carioca_-2+quismo,0! carnal_0,-2+e!
caro_!-1+estia,0! carol_+ice,+a! caron_0,+a! carpal_0,-2+o! carrea_+cao,+dor! carret_+o,0! casamentei_0,-2+o!
casar_-1+mento,-1+dor! casca_-5+descascamento,-5+descascador! caseir_0,-3+a! casmurr_+ice,+o!
cassa_+cao,+dor! castic_+ismo,+o! castiga_-1+o,+nte! casto_-1+idade,0! castra_+cao,+dor! cataloga_-1+o,+dor!
catar_!-1,-1+dor! catartic_0,-3+se! catastrofi_0,-1+e! catedratic_0,-3! catequiza_-3+ese,+dor! catingu_0,-1+a!
cativ_-5+seducao,+ante! catolic_+ismo,+o! caturr_+ice,+a! caudalos_-2,0! caulinar_0,-4+e! causal_0,-1!
causar_-1,-1+dor! cautel_+a,oso! cear_!-2+ia,-1+dor! cece_+io,+ador! cedular_0,-1! cegar_!-2+ueira,-1+nte!
cego_-1+ueira,0! celebra_+cao,+dor! celebre_-1+idade,0! celest_0,-4+u! celesti_0,-5+u! cels_+itude,+o!
celular_0,0! celular_0,-1! cemiterial_0,-2+o! cenic_0,-2+a! censual_0,-2+o! centrifug_+acao,+dor!
centrossomatic_-1+a,+o! centurial_0,-1! ceptic_-4+ticismo,+o! cerca_-0,+nte! cerebral_0,+o! cerimonia_0,0!
cerquei_-4+ca,+ro! certeir_-7+precisao,+o! cerval_0,-2+o! cerz_+idura,+idor! cesare_0,-1! cessar_-1+o,-1+dor!
cetic_+ismo,+o! ceva_!-0,+dor! chacin_+a,+ador! chacoalha_-9+chateacao,+dor! chafurd_+ice,+ador!
chamar_-1+mento,-1+dor! chamativ_0,0! chamb_+oice,+ao! chamus_+camento,+cador! chancel_+a,+er!
chapot_+a,+ador! charlata_+nice,+o! charmos_-2+e,0! chatea_+cao,+dor! chato_-1+ice,0! chavasc_-1+quice,+o!
checa_+gem,+dor! chefiar_-1+,-3+e! chega_+da,+dor! cheira_-1+o,+dor! cheiro_-0,+so! chelosvaco_0,-2+quia!
chiar_-1+da,-1+dor! chiban_+ca,+te! chibat_+ada,+ador! chilen_0,-1! chinc_+ada,+ador! chiqu_+ismo,+e!
choca_-2+que,+nte! chocalh_+ice,+eiro! choch_+ice,+o! chor_+o,+ao! chove_-3+uva,+dor! chule_+io,+ador!
chulo_-1+ice,0! chumba_+gem,+dor! chup_+ada,+ador! churrasca_+da,-2+queiro! chuta_-1+e,+dor!
chuvos_-2+a,0! ciar_!-1+da,-1+dor! cibernetic_0,+a! ciclic_0,-2+o! ciclonic_0,-2+e! ciclopic_0,-2+e!
ciclotimic_-1+a,+o! cidada_+nia,+o! cientific_0,-5+cia! ciliar_!-2+os,0! cincar_-1+da,-1+dor! cinderel_0,+a!
cinic_-1+smo,+o! cioso_0,0! circular_0,-2+o! circunspe_+ccao,+cto! cismar_-1,-1+dor! citadin_0,-5+dade!
citav_-1+cao,0! citre_-5+acidez,+o! citri_-5+acidez,0! cium_+e,+ento! civel_!0,0! civico_-2+smo,0!
civil_!0,-3+dadao! clamar_-2+or,-1+dor! clandestin_+idade,+o! claro_-1+eza,0! claustral_0,-2+o! clausular_0,-1!
clavicular_0,-1! clerical_0,-2+o! clicar_-3+que,-1+dor! clinic_?0?+a,?+a?+o! clinica_-0,0! cliva_+gem,+dor!
clown_+ismo,-1+n! coagment_+o,+ador! coalh_+ada,+ador! coaxial_0,-7+eixo! cobaltiz_+agem,+ador!
cobert_+ura,+o! cobica_-0,+dor! cobrar_-1+nca,-1+dor! cobri_-2+ertura,+dor! cocar_!-2+eira,-1+dor!
coceg_+as,+uento! cochar_-1,-1+dor! cochich_+o,+ador! cochil_+ada,+ador! coclear_0,-1! coercitiv_-4+ao,+o!
cognatic_-2+cao,+o! coitado_-7+desgraca,0! colar_!-1+gem,-1+dor! colegia_+tura,+do! coleric_0,-2+a!
coleta_-0,+dor! colhe_+ita,+dor! colid_-1+sao,+ente! coligar_-1+cao,-1+dor! colmat_+agem,+ador!
coloquial_0,-2+o! colossal_0,0! colunar_0,-1! comanda_-1+o,+nte! combat_?+ividade?+e,+ente! combustive_0,0!
comeca_-1+o,+dor! comedi_+mento,+do! comenta_+rio,+dor! comer_!-2+ida,-1+dor! comerci_+o,+ante!
comercial_0,-2+o! cometari_0,-2! comil_+anca,+ao! cominu_+icao,+inte! comissural_0,-1! compac_+idade,+to!
companheir_+ismo,+o! comparav_-1+cao,0! compassa_-1+o,+dor! compensa_+cao,+dor! compet_+icao,+idor!
competen_+cia,+te! complementa_?+ridade?+cao,+dor! comportar_-1+mento,-1+dor! composi_+cao,+tor!
compost_+ura,+o! comprar_-1,-1+dor! compraze_-8+prazer,-8+prazeroso!
compreen_?+sividade?+sao,-8+entendedor! compreensiv_-1+bilidade,0! comprid_-1+mento,0!
comprobat_-3+vacao,0! compun_+cao,+gente! comung_-1+hao,+ante! comunica_+cao,+dor! concav_+idade,+o!
conceb_-1+pcao,+edor! concede_-2+ssao,+nte! concentra_+cao,+dor! concentric_+idade,+o!
concern_+encia,+ente! concert_+o,+ista! concessi_-1+ao,0! concho_-3+fianca,0! concomit_+ancia,+ante!
concorda_+ncia,+nte! concreta_+gem,+dor! concreto_-1+ude,0! condal_0,-2+e! condensa_+cao,+dor!
condiz_-6+adequacao,+ente! condominial_0,-2+io! conferen_+cia,+te! confessa_-4+issao,-1+o! confia_+nca,+do!
confiav_-1+nca,0! confiden_+cia,+te! confisc_+o,+ador! conflit_+o,+ante! conformista_-2+mo,0!
confort_+o,+ador! confortav_0,-2+o! confront_+o,+ador! congenia_0,-8+genio! congressual_0,-2+o!
conjectur_+a,+ador! conjetur_-3+ctura,-3+cturador! conjugal_0,-2+e! conjumin_+ancia,+ante! conjuntival_0,-2+o!
conjunto_-3+gacao,0! conjuntural_0,-1! consciencial_0,-1! conseg_-1+cucao,+uidor! consert_+o,+ador!
considerav_-1+cao,0! consola_+cao,+dor! constar_!-2+ituicao,-1+nte! constata_+cao,+dor!
constitucional_0,-6+icao! constru_+cao,+tor! construtivis_+mo,+ta! consubstancia_+cao,+dor! consular_0,-2+e!
consult_?0?+a,?+a?+ador! consum_+o,+idor! contact_-2+to,+ante! contagia_-1+o,+nte! contar_!-1,-1+dor!

contat_+o,+atante! conteir_+a,+ador! contempla_+cao,+dor! conten_+cao,+dor! content_+amento,+ador!
contesta_+cao,+dor! continental_0,-2+e! contra_0,0! contrabalanca_-1+o,+dor! contrabandea_-2+o,-2+ista!
contracambi_+o,+ante! contracena_-0,+dor! contracol_+agem,+ador! contradi_+cao,+torio! contrari_+edade,+o!
contraria_-1+edade,+dor! contrasta_-1+e,+nte! contratual_0,-3+o! control_+e,+ador! controvers_+ia,0!
convenia_-1+o,+dor! conversa_-0,+dor! convic_+cao,+to! convida_-2+te,+dor! convinc_-3+encimento,-3+encedor!
convir_-2+eniencia,-2+eniente! convolu_+cao,+cionador! copar_!-1+gem,-1+dor! copej_+ada,+ador!
copia_-0,+dor! copios_-6+abundancia,0! copular_-1,-1+dor! coquet_+ismo,+e! coral_0,0! corar_-1,-2+ador!
corcov_+adura,+ador! cordat_-2+ura,+o! corintian_0,-3+hians! corneo_0,-2+o! coroar_-1+cao,-1+dor!
corporal_0,-3! corporativis_+mo,+ta! corporativo_-1+smo,0! corpore_0,-2! correla_+cao,+cionador!
corrente_!-8+ocorrencia,0! correr_-2+ida,-1+dor! correta_+gem,-1+or! corretiv_-3+cao,+o! correto_-2+cao,0!
corrig_-2+ecao,-2+etor! corriqueir_+ice,0! corroe_-1+sao,+nte! corromp_-3+upcao,-3+uptor! cortical_0,-2+e!
coser_!-1+dura,-1+dor! cosmic_0,-2+o! costal_0,-2+ela! costea_+gem,+dor! costeir_0,-3+a! costum_+e,+eiro!
cota_+cao,+dor! cotiar_-6+cotidiano,-1+do! cotidian_0,+o! coxal_0,-1! coz_+imento,+inheiro!
cozinha_-3+mento,-1+eiro! crass_+idao,+o! credita_-1+o,+nte! creditori_0,-2! crendeir_+ice,+eiro!
cresp_+idao,-5+encrespador! crest_+a,+ador! cretace_0,+o! cretin_+ice,+o! criar_-1+cao,-1+dor!
criativ_+idade,+o! criminal_0,-4+e! crimino_-1+alidade,+so! crista_+ndade,+o! cristalin_+idade,+o!
cristaliz_+acao,+ador! critic_+a,+o! criticav_-1,0! crocante_0,0! croma_+gem,+dor! cromatic_0,-7+or!
cromossomic_0,-2+o! cronometr_+agem,+ador! cru_!+eza,0! crucial_0,0! cruel_+dade,0! cruento_-3+za,0!
crural_0,-5+oxa! cubar_-1+gem,-1+dor! cubital_0,-2+o! cubr_-3+obertura,-3+obridor! cucurbital_0,-2+o!
cuida_+do,+doso! culinari_0,+a! culp_+a,+ador! cultiv_+o,+ador! cultural_0,-1! cumplic_+idade,+e!
cumpriment_+o,+ador! cupular_-1,-1+dor! curav_-1,0! curial_0,-1! curra_-0,+dor! curricula_0,-1+o!
curto_-1+eza,0! curva_-0,+dor! cusp_+idura,+idor! cuspinh_+adura,+ador! custa_-1+o,-1+oso! custea_-1+io,+dor!
custo_-0,+so! cutaneo_0,-4+is! cuticular_0,-1! cutuc_+ada,+ador! danad_-1+cao,-1+do! danar_!-2+o,-1+do!
danca_-0,+rino! dancav_0,-1! dand_+ismo,+e! danifica_-5+o,+dor! daninh_+eza,+o! dar_-2+oacao,-2+oador!
datar_!-1+cao,-1+dor! deband_+ada,+ador! debit_+o,+ador! debocha_-1+e,+do! debulha_+cao,+dor!
debut_+e,+ador! decalc_-1+que,+ador! decamp_+amento,+ador! decan_+ato,+o! decap_+agem,+ador!
decep_+amento,+ador! decid_-1+sao,-1+sor! decisiv_-2+ao,0! decola_+gem,+dor! decomposi_+cao,+tor!
decre_+scimo,+mentador! decrept_+ude,+o! decresce_-1+imo,+nte! decreta_-1+o,+dor! decup_+agem,+ador!
dedar_-2+uragem,-2+urador! dedur_+agem,+ador! defasa_+gem,+dor! defeca_+cao,+dor! defeit_+o,0!
defende_-3+sa,-2+sor! defensav_0,-4+sa! defensiv_0,-4+sa! deficit_+e,+ario! defle_+xao,+ctor!
defunto_-7+morte,0! degel_+o,+ador! deglut_+icao,+idor! degolar_-1,-1+dor! degrada_+cao,+dor!
deita_+da,+dor! deix_+a,0! delata_-2+cao,-1+or! deleita_-1+e,+dor! delgad_-6+fino,+o! delica_+deza,0!
delicad_+eza,+o! delici_+a,+ador! delicios_-2+a,0! delinqu_+encia,+ente! delir_+io,+ante! demanda_-0,+nte!
demasiad_-1,0! demit_-1+ssao,+idor! demofil_+ia,0! demoniac_0,-2+o! demonstra_+cao,+dor! demora_-1+a,+do!
dendroclast_+ia,+a! dengo_-1+uice,+so! denod_+o,0! denso_-1+idade,0! dentari_0,-3+e! denu_+ncia,+nciante!
denuncia_-0,+nte! dependur_+a,+ador! deplorav_-1+cao,0! depreda_+cao,+dor! depreend_-1+sao,-1+sor!
depress_-7+pressa,+ivo! dermopapilar_0,-1! derrama_+mento,+dor! derrap_+agem,+ador! derreng_+ue,+ador!
derrete_-1+imento,+dor! derrib_+amento,+ador! derrot_+a,+ador! derrub_+amento,+ador! desabaf_+o,+ador!
desabala_-8+precipitacao,0! desabar_-1+mento,-1+dor! desabon_+o,+ador! desabrig_+o,+dor!
desabus_+o,+ador! desacat_+o,+ador! desacopl_+amento,+ador! desafi_+o,+ador! desafog_+o,+ador!
desafora_-1+o,+do! desafort_-8+infelicidade,0! desafront_+a,+ador! desagrada_-1+o,+dor! desagrav_+o,+ador!
desajust_+e,+ador! desampar_+o,+ador! desanda_-0,+nte! desanima_-1+o,+dor! desanoj_+o,+ador!
desanuvi_-8+serenidade,+ador! desarmar_-1+mento,-1+dor! desassosseg_+o,0! desastra_-1+e,+do!
desastros_-2+e,0! desatrel_+agem,+ador! desavez_+o,+ador! desbloque_+io,+ador! desbocado_0,0!
desbord_+o,+ante! desbund_+e,+ante! descaminh_+o,+ante! descans_+o,+ador! descart_?0?+e,?+e?+ador!
descart_+e,+ador! descasc_+amento,+ador! descer_!-2+ida,-1+dor! descerc_+o,+ador! descol_+agem,+ador!
descompassa_-1+o,+dor! descompost_+ura,+o! desconfia_+nca,+do! desconfortav_0,-2+o! desconta_-1+o,+dor!
descontent_+amento,+ador! descontrol_+e,+ador! descorn_+amento,+ador! descortin_+o,+ador!
descos_+edura,+idor! descuida_+do,+doso! desculp_+a,+ador! descultiv_+o,+ador! desde_+m,+nhador!
desdenh_-2+m,+ador! desdize_-7+contradicao,-7+contraditorio! desdour_+o,+ador! desejav_0,-2+o!
deselegant_-1+cia,0! desembarg_+o,+ador! desempach_+o,+ador! desempat_+e,+ador! desempeg_+o,+ador!
desempen_+o,+ador! desempenh_+o,+ador! desencaix_+e,+ador! desencalh_+e,+ador! desenfard_+o,+ador!
desengan_+o,+ador! desengasg_+ue,+ador! desengat_+e,+ador! desengonc_+o,0! desengross_+o,+ador!
desenha_-1+o,-1+ista! desenle_+io,+ador! desenred_+o,+ador! desensin_+o,+ador! desentros_+amento,+ador!
desentulh_+o,+ador! desenvolt_+ura,+o! desenxofr_+amento,+ador! desenxovalh_+o,+ador! deserda_+cao,+nte!
deserta_-2+cao,-1+or! desertic_0,-2+o! desesper_+o,+ado! desestiv_+a,+ador! desfadig_+a,+ador!
desfalc_-1+que,+ador! desfaze_-7+destruicao,-7+destruidor! desfecha_-1+o,+dor! desferi_-4+fecho,+dor!
desfia_+dura,+dor! desfil_+e,+ador! desfrut_+e,+ador! desfrutav_0,-2+e! desgarr_+e,+ador! desgast_+e,+ador!
desgraca_-0,+nte! desgrud_-7+afastamento,+ador! desguia_-7+abandono,+dor! desilu_+sao,+sor!
desinencial_0,-1! desinteress_+e,+ante! deslaje_+amento,+ador! desliza_+mento,+dor! deslustr_+e,+ador!
desmaia_-1+o,+dor! desmama_-1+e,+dor! desmanch_+o,+ador! desmand_+o,+ador! desmembra_+mento,+dor!
desment_+ido,+idor! desmoit_+a,+ador! desmold_+agem,+ador! desnud_-6+nudez,+ador! desobstr_+ucao,+uidor!
desolar_-1+mento,-1+dor! desorde_+m,+iro! desov_+a,+ador! despach_+o,+ador! despedi_+da,+dor!
despel_+a,+ador! despend_-2+sa,+edor! desperdic_+io,+ador! desperta_+r,+dor! despi_-5+nudez,+do!
despist_+e,+ador! despolu_+icao,+ente! despossuid_0,0! desprende_-1+imento,+dor! despreza_-1+o,+dor!
despronu_+ncia,+nte! desram_+a,+ador! destaca_-2+que,+dor! desterr_+o,+ador! destina_-1+o,+dor!
destoa_+mento,+nte! destrinc_+a,+ador! destro_!-1+eza,0! destroc_+o,+ador! destru_+icao,+idor!
desvel_+o,+ador! desvirtu_+amento,+ador! deter_!-1+ncao,-1+ntor! determinis_+mo,0! deterr_+encia,+ente!
detesta_-7+odio,+dor! detetives_0,-1! detona_+cao,+dor! devagar_-7+lentidao,0! devanea_-1+io,+dor!
devassa_+mento,+dor! devasso_-1+idao,0! devasta_+cao,+dor! dever_-4+ivida,-1+dor! devido_0,-3+er!
devora_+cao,+dor! devota_-2+cao,-1+o! diabetic_0,-2+e! diabol_0,-1! diaconal_0,-2+o! diacronic_-1+a,+o!
diafragmatic_-1+a,+o! diagnostica_-1+o,+dor! diagnostico_-4+e,0! dialetal_0,-2+o! dialetic_0,+a!
dialoga_-1+o,+dor! dianteir_0,+a! diario_0,0! didatic_+a,+o! diferen_?+ca?+ciacao,+ciador!
diferi_-1+enca,-1+encador! dific_+uldade,+ultador! difu_+sao,+sor! digeriv_-3+stao,0! digital_!0,-2+o!
digladia_-8+gladio,+dor! dign_+idade,+o! dilata_+cao,+dor! diletant_+ismo,+e! dilu_+icao,+ente!
diminu_+icao,+idor! dinamic_-1+smo,+o! dioarreic_0,-1+a! dioces_0,+e! dioxinic_0,-2+a!

diplomar_-1+cao,-1+dor! diplomatic_0,-3+cia! diretorial_0,-2+o! dirig_-2+ecao,+ente! dirij_-2+ecao,-1+gidor!
dirimi_-6+eliminacao,-6+eliminador! disca_+gem,+dor! discerniv_-1+mento,0! disciplin_+a,+ador!
discorda_+ncia,+nte! discrepa_+ncia,+nte! discret_-2+icao,+o! discri_+cao,+tor! discricionari_+edade,+o!
discursiv_0,-2+o! discut_-1+ssao,+idor! discutiv_-3+ssao,0! disfarca_-1+e,+dor! disle_+xia,0! dispar_+idade,0!
dispara_-1+o,+dor! disparat_+e,+ador! dispend_+io,+ioso! dispendios_-1,0! dispensa_-0,+dor! dispensav_-1,0!
disputa_-0,+dor! disseca_+cao,+dor! dissipa_+cao,+dor! disson_+ancia,+ante! disting_-1+cao,+uidor!
distrital_0,-2+o! ditar_-1+do,-1+dor! ditatorial_0,-6+dor! diurn_0,-3+a! divin_+dade,+o! divulga_+cao,+dor!
dize_-4+declaracao,-4+declarante! doad_-1+cao,+or! doar_-1+cao,-1+dor! dobl_+ez,+e! dobrav_0,-1!
doce_-1+ura,0! documentari_0,-3+o! doenti_-2+ca,+o! doer_!-2+r,-1+nte! dogal_0,-2+e! dolor_0,-3+r!
dolos_-1,0! dom_+a,+ador! domar_-1,-1+dor! domestica_+cao,+dor! domestico_0,-9+casa! domiciliar_0,-2+o!
domingu_0,-1+o! dominical_!0,-4+go! doravant_-8+futuro,0! dormi_+da,+dor! dorsal_0,-2+o! dosar_-2+e,-1+dor!
dota_+cao,+dor! doutoran_-1+do,0! drastic_-7+rigor,0! dribl_+e,+ador! drogar_-1,-1+dor! dubio_-1+edade,0!
dubla_+gem,+dor! ducal_0,-3+que! duelar_-2+o,-2+ista! duodenal_0,-2+o! duplo_-1+icidade,0!
dura_+cao,+douro! duradour_-4+cao,+o! durao_-2+eza,0! durav_-1+cao,0! duro_!-1+eza,0! duvid_+a,-5+cetico!
eclesiast_0,-9+igreja! eclips_+e,+ante! ecoa_-1+,-1! economiz_-1+a,+ador! ecumenic_-1+smo,+o! edenic_0,-2!
editor_?0?+acao,-1+r! educa_+cao,+dor! educacion_0,-3+ao! edulc_-5+docura,+orante! efun_-1+sao,-1+sor!
egipci_0,-3+to! ejet_+amento,+ador! elegant_-1+cia,0! elege_-2+icao,-2+itor! eleitoral_0,-5+cao!
eleitorei_0,-5+cao! elementar_+idade,0! elenca_-1+o,+dor! eletrocuta_-2+cao,-1+or! elid_-1+sao,-1+sor!
elogi_+o,+ador! elucubra_+cao,+dor! emanar_!-1+cao,-1+dor! embala_+gem,+dor! embals_+amento,+ador!
embasbac_-8+espanto,+ado! embebeda_-8+bebedeira,-5+riagador! embolar_-1+cao,-1+dor!
embrac_+adura,+ador! embrion_0,-2+ao! embuc_+adela,+ador! embut_+idura,+idor! emend_+a,+ador!
empach_+a,+ador! empalma_-7+palma,+dor! empana_+da,+dor! empapuc_+amento,+ador!
emparv_+amento,+ador! empat_+e,+ador! empenh_+o,+ador! empesg_+adura,+ador! empin_+o,+ador!
empiriocritic_+ismo,+o! empoeir_-7+poeira,+ador! empolei_-7+poleiro,+rador! empolga_+cao,+nte!
emposs_-6+posse,+ador! empreende_-1+imento,+dor! empregati_0,-3+o! empreit_+ada,+eiro! empresar_-1,+io!
empresarial_0,-1! empunha_+dura,+dor! enamora_-7+namoro,-7+namorador! encabel_+adura,+ador!
encaix_+e,+ador! encalh_+e,+ador! encali_+dela,+dor! encampa_+cao,+dor! encano_+amento,+ador!
encanud_+amento,+ador! encapsula_+mento,+dor! encarce_-7+carcere,+rador! encard_+imento,+idor!
encarg_+o,+ador! encarnica_+mento,+dor! encaroc_+ada,+ador! encarrega_-1+o,+dor! encasc_-1+que,+ador!
encena_+cao,+dor! encesta_-7+cesta,+dor! encomenda_-0,+dor! encorp_+adura,+ador! encrenca_-0,+dor!
encresp_+amento,+ador! encru_+amento,+ador! encuca_+cao,+dor! encurrala_+mento,+dor!
encurv_+amento,+ador! endeusa_+mento,+dor! endiabra_+gem,+do! endireita_-9+correcao,+dor!
endotelial_0,-2+o! energetic_0,-4+ia! enfad_+o,+onho! enfaix_+e,+ador! enfar_+o,+ador! enfarinh_+adela,+ador!
enfasti_+amento,+ador! enfatic_-3+se,+o! enfeiar_-1+mento,-1+dor! enfeita_-1+e,+dor! enferm_+idade,+o!
enfilei_-7+fila,+rador! enform_+agem,+ador! enfrenta_+mento,+dor! engai_+o,+ador! engaiola_-8+gaiola,+dor!
engambela_+cao,+dor! engan_+o,+ador! engasg_+ue,+ador! engast_+e,+ador! engavet_+amento,+ador!
engazop_+amento,+ador! engenha_-1+o,+dor! engloba_+mento,+dor! engod_+amento,+ador!
engoma_+gem,+dor! engord_+a,+ador! engordura_+mento,+dor! engrad_+amento,+ador!
engranz_+amento,+ador! engravi_-7+gravidez,+dante! engraxa_+da,+dor! engul_-2+olicao,-2+olidor!
enjoa_-1+o,+do! enlac_+e,-1+cador! enlame_+adura,+ador! enlea_-1+io,-5+enrolador! enlous_+amento,+ador!
enlut_-5+luto,0! enrasc_+ada,+ador! enred_+o,+ador! enros_+camento,+cador! enrosca_+mento,+dor!
ensabo_+ada,+ador! ensaia_-1+o,+dor! ensambl_+adura,+ador! ensarilha_-1+o,+dor! ensil_+agem,+ador!
ensimesm_+amento,+ador! ensin_+o,+ador! ensolara_0,-8+sol! ensopa_+do,_+dor! entabu_+amento,+ador!
entardec_+er,+edor! entedia_-7+tedio,+nte! enterra_-1+o,+dor! entoa_+da,+dor! entorn_+adura,+ador!
entort_+adura,+ador! entranha_-8+penetracao,+dor! entreg_+a,+ador! entrega_-0,+dor! entremea_-1+io,+dor!
entreolha_+do,+dor! entreouvi_-9+audicao,+nte! entressonh_+o,+ador! entrete_+nimento,+dor!
entrich_-7+trincheira,+eirador! entristec_-9+tristeza,+edor! entros_+amento,+ador! entusias_+mo,+ta!
enuncia_+cao,+dor! envelop_+amento,+ador! enverga_+dura,+dor! envergonh_-9+vergonha,+ador!
enxagu_+e,+ador! enxamea_-1,+dor! enxert_+o,+ador! enxofr_+amento,+ador! epico_!0,0! episcopal_0,-1+do!
episodi_0,+o! epistolar_0,-1! epitelial_0,-2+o! equatorial_0,-6+dor! equestre_0,-8+cavalo! equilater_0,+o!
equilibr_+io,+ista! equino_0,-6+cavalo! equiparav_-1+cao,0! equiponder_+ancia,+ante! equivoc_+o,+ado!
eral_0,-2+o! eret_-1+cao,+o! erica_+mento,+do! erigi_3+guimento,-3+etor! ermar_-1+mento,-1+dor!
ermo_-4+solidao,0! erradicav_-1+cao,0! errar_!-2+o,-1+do! erratic_0,+o! errone_-2,0! erudi_+cao,+to!
erv_+agem,+ador! esbarro_+camento,+cador! esbelt_-6+beleza,+o! esboc_+o,+ador! esbof_+amento,+ador!
esborr_+o,+ador! esborralh_+ada,+ador! esborrat_+adela,+ador! escabel_+a,+ador! escalavr_+amento,+ador!
escald_+a,+ador! escamotea_+cao,+dor! escancar_+o,+ador! escandalos_-1,0! escandi_-7+decomposicao,+dor!
escap_+e,+ador! escapular_0,-1! escass_+ez,+o! escleros_+e,+ado! escolh_+a,+edor!
esconde_-7+ocultamento,+dor! escora_+mento,+dor! escorrer_-2+imento,-1+dor! escorrid_-2+encia,+o!
escorropich_+adela,+ador! escorv_+amento,+ador! escrach_+o,+ador! escrav_+idao,+izador!
escrev_-2+ita,-2+itor! escrotal_0,-2+o! escrupulos_-1,0! esculach_+o,+ador! esculpi_-2+tura,-2+tor!
escult_0,+ura! escuma_-0,+dor! escuro_!-1+idao,0! escus_+a,+ador! escuta_-0,+dor! esfalf_+amento,+ador!
esfenoidal_0,-2+e! esferic_+idade,+o! esfigm_+ismo,+o! esfoaca_+mento,+nte! esfola_+mento,+dor!
esforca_-1+o,+do! esfreg_+ada,+ador! esgalh_+a,+ador! esgar_!-3+carnio,-1+dor! esgarca_+dura,+dor!
esgrim_+a,+ista! esgui_+ez,+o! esguich_+o,+ador! eslovaco_0,-2+quia! esmalt_+agem,+ador!
esmar_-5+avaliacao,-5+avaliador! esmera_-1+o,+do! esmol_+aria,+er! esnob_+acao,+e! esnob_+icidade,+ador!
espacial_0,-3+o! espalma_+da,+dor! espanhol_0,-2+a! espanta_-1+o,+dor! esparram_+e,+ador!
espartan_+idade,+o! especia_0,0! especific_?+idade?+acao,+ador! esper_+a,+ador! esperne_+io,+ador!
esperta_+mento,+dor! esperto_-1+eza,0! espesso_-1+ura,0! espetacular_0,-2+o! espetar_-1+da,-1+dor!
espevita_+mento,+dor! espiar_-1+da,-1+o! espinc_+a,+ador! espinhal_0,-1! espion_+agem,-2+ao! espiral_0,0!
esplend_+or,+ente! esponsal_0,-2+o! espraia_+mento,+dor! espreit_+ada,+ador! esprem_+edura,+edor!
espuma_-0,+do! espur_+iedade,0! esquec_+imento,+ido! esqueletic_0,0! esquisit_+ice,+o! esquivo_-1+ez,0!
esquizofrenic_-1+a,+o! estaciona_+mento,+dor! estacionari_0,+o! estadual_0,-3+o! estafa_-1+a,+nte!
estamp_+a,+ador! estanh_+agem,+ador! estar_!-1+do,0! estatal_0,-3+do! estatic_0,+o! estatistic_0,+a!
estatutar_0,-2+o! estelar_0,-4+rela! estende_-7+extensao,+dor! esternal_0,-2+o! estesiogen_+ia,+ico!
estetiza_-2+ca,-2+cista! estiar_-1+gem,-1+dor! estica_+mento,+dor! estilar_0,-2+ete! estima_-6+apreciacao,+dor!
estimul_+o,+ante! estof_+amento,+ador! estomacal_0,-3+go! estomatic_0,-2+o! estour_+o,+ador!

estouv_+amento,+ador! estrafeg_+o,+ador! estrag_+o,+ador! estrangula_+mento,+dor! estranh_+eza,+ador!
estrea_-1+ia,+nte! estreiar_-1,-1+nte! estreit_?+eza?+amento,+ador! estrelar_0,-1! estremar_-1+dela,-1+edor!
estremunha_+mento,+dor! estressa_-1+e,+nte! estria_+mento,+dor! estrit_-6+restricao,+o! estuda_-1+o,+nte!
estudant_0,+e! estudios_-3+o,0! estult_+ice,0! estupefa_+cao,+to! estupor_-0,+ador! estupr_+o,+ador!
esvai_-1+ecimento,-1+ecedor! esvazi_+amento,+ador! esverdea_-8+verde,+nte! etario_0,-6+idade! etere_0,-1!
eterno_-1+idade,0! etico_0,-1+a! etmoidal_0,-2+o! eucaristic_+a,+o! eufratic_0,-2+es! eugene_+sia,0!
eumatic_0,-1+a! eustatic_-3+sia,+o! eutireoid_+ismo,+e! evasiv_-2+ao,0! evenc_-3+iccao,-3+ictor!
evict_-1+cao,+o! eviden_+cia,+ciador! eviden_+cia,+te! evolui_-1+cao,+dor! exagera_-1+o,+do! exala_+cao,+dor!
exalta_+cao,+dor! exam_+e,+inador! exat_+idao,+o! exato_-1+idao,0! exauri_-2+stao,-2+stor! exaust_+ao,+ador!
exaustivo_-3+ao,0! excede_-2+sso,+nte! excels_+itude,+o! excentric_+idade,+o! excepciona_-6+cao,0!
excessiv_-2+o,+o! excet_-1+cao,+uador! exclusiv_+idade,+o! excret_-1+cao,0! executa_-2+cao,-1+or!
executav_-1+cao,0! exerc_+icio,+itante! exerga_-6+visao,+dor! exilar_-2+io,-1+dor!
eximir_-6+isencao,-6+isentador! exorta_+cao,+dor! exotic_-1+smo,+o! expand_-1+sao,-1+sor!
expedicion_0,-3+ao! experien_+cia,+te! explicit_+acao,+ador! expor_!-1+sicao,-1+sitor! expressa_+o,+dor!
expulsa_+o,+dor! expung_-1+cao,+ador! expurga_-1+o,+dor! exsolv_-1+ucao,+edor! exspu_+icao,+idor!
exsucta_-2+cao,-1+or! extas_+e,+iador! extasia_-2+e,+dor! extermin_+io,+ador! externa_-7+declaracao,+dor!
externo_0,-2+ior! exting_-1+cao,+uidor! extint_-1+cao,+o! extorq_-1+sao,-1+sor! extra_0,0!
extravag_+ancia,+ante! extrem_+idade,+ador! fabrica_+cao,+nte! fabril_0,-1+ca! fabulos_-2+a,0!
faccios_+ismo,+o! face_+cia,+to! faceir_+ice,+o! facial_0,-3+e! facular_0,-1! facult_-1+dade,+ivo!
fageden_+ismo,+ico! fala_-0,+nte! falangeal_0,-2! falec_+imento,+ido! falho_-1+a,0! fali_-1+encia,+do!
falic_0,-2+o! fals_+idade,+o! falt_+a,+oso! faminto_-6+ome,0! famoso_-3+a,0! fanchon_+ice,+o!
fanfarr_+ice,+eiro! fanhos_0,0! fantasia_-0,+dor! fantastic_0,0! faraonic_0,-3! farelace_0,-3+o! farfalh_+o,+ador!
farinace_0,-3+ha! farrea_-2+a,+dor! farsesc_0,-3+a! farta_-1+ura,+dor! fatiar_-1,-1+dor! fatidic_-7+tragedia,+o!
fatiga_-0,+nte! fatuo_-1+idade,-1+o! favorit_+ismo,+o! fazendari_0,-2! fazer_!-5+construcao,-5+construtor!
febril_0,-2+e! fecal_!0,-3fezes! fede_-1+or,-1+orento! federal_0,-1+cao! fedorent_-3,+o! feeric_0,-5+ada!
feio_-1+ura,0! feitic_+o,+eiro! feldspatic_0,-2+o! femini_+lidade,0! fende_-4+issao,+dor! feri_+mento,+dor!
fero_+cidade,0! ferra_+gem,+dor! ferre_0,-1+o! ferrenh_0,-3+o! ferv_+ura,+edor! fest_0,+a! fetid_-3+dor,0!
feudal_0,-2+o! fiar_-1+nca,-4+afiancador! fibros_-2+a,0! ficar_-5+permanencia,0! fictici_-4+cao,0! fidalg_+uia,+o!
fidalgu_+ice,+o! filarmoni_0,-9+musica! filet_+agem,+ador! filhar_-1+cao,-6+pai! filho_-1+acao,0! filial_0,0!
filma_+gem,+dor! filmic_0,-2+e! filoarmonic_0,+a! filogenetic_0,-4+ia! filosof_+ia,+o! final_0,0!
financ_?0?+iamento,?+a?+idador! finar_-1+mento,-1+do! finda_-5+finaliza??o,+dor! fino_-1+ura,0! firme_+za,0!
fiscal_0,-2+o! fisga_+da,+dor! fitar_-5+visao,-1+dor! fixo_-1+idez,0! flagra_+nte,+dor! flam_+ancia,+ante!
flamba_+gem,+dor! flamenguist_0,-4+o! flecha_+da,+dor! flerta_-1+e,+dor! flexiona_-4+ao,+dor! flexo_-1+ao,0!
floral_0,-2! florea_-1+io,-2+ista! florestal_0,-1! florir_-2+acao,-2! floristic_0,-5+a! fluvial_0,-7+rio!
focar_-1+mento,-1+dor! foder_!-2+a,-1+dor! fofar_!-5+afofamento,-1+dor! fofo_!+ura,0! fofoca_-0,-2+queiro!
folcloric_0,-2+e! folgar_-1,-1+do! folhar_-1+da,-1+dor! folhea_+da,+dor! folhetin_0,-1+m! foliace_-4+lhas,+o!
foment_+o,+ador! foragi_-5+uga,-5+ugitivo! forc_+a,+ador! formic_0,+o! formicular_0,-5+ga! formid_0,0!
formos_+ura,+o! forro_!-5+alforria,0! fortal_+eza,+ecedor! forte_-2+ca,0! fosfatic_0,-2+o! fossa_+dura,+dor!
fossil_0,0! fotogen_+ia,+ico! fotostatic_-2+sia,+o! fracassa_-1+o,+do! fraco_-2+queza,0! franc_-1+queza,+o!
frances_0,-2+a! francofil_+ia,0! franquia_-0,+dor! frasal_0,-2+e! frascar_+ia,+io! frase_+io,+ador!
fratur_+a,+ador! fraud_+e,+ador! frem_+ito,+ente! frenetic_0,0! fresc_+ura,+o! fret_+amento,+ador!
frigi_-2+tura,-2+tador! frigorific_+acao,+o! frio_!-1+eza,0! fris_+agem,+ador! frita_-1+ura,+dor! frivol_+idade,+o!
fronteir_0,+a! froux_+idao,+o! fru_+icao,+idor! frutar_!-2+ificacao,-2+ificador! frutif_0,-2+a! fugi_-1+a,+tivo!
fugid_0,-2+a! fulgu_-1+or,+rante! fulvid_0,0! fumar_-2+o,-1+nte! funciona_+mento,0! funda_+cao,+dor!
fundiari_0,+o! fundir_-4+sao,-1+dor! fundo_-5+profundidade,0! funeb_0,-5+morte! funerari_0,+a!
furibundo_-5+a,0! furios_-2+a,0! furt_+o,-4+ladrao! futeb_0,+ol! futebolistic_0,-5! futura_+cao,+dor! futuro_-0,0!
gago_!-1+ueira,0! gaguej_-1+ira,-3+o! gala_+nice,-1! galactic_0,-4+xia! galant_+aria,+eador! galhard_+ia,0!
galop_+e,+ante! galre_+io,+ador! galvanotip_+agem,+ador! gamar_-1+cao,-1+do! gamenh_+ice,+o!
gamgren_+a,+ador! ganancios_-2+a,0! ganglionar_0,-3! gangster_+ismo,0! ganhar_-2+o,-1+dor!
garant_+ia,+idor! garfar_-1+da,-1+dor! gargalha_+da,+dor! garimp_+o,+eiro! garrafal_0,0! garrul_+ice,+o!
gasoso_-3,0! gasta_-1+o,+dor! gastric_0,-7+estomago! gastrolatr_+ia,a! gazetari_0,-3+a! gear_!-1+da,-1+da!
gelar_-2+o,-1+dor! gelid_0,-2+o! gemar_-1,-1+dor! gemer_-2+ido,-1+dor! generic_0,+o! genetic_0,+a!
gengival_0,-1! genital_0,0! gentil_+eza,0! genuin_+idade,+o! geral_-3+neralidade,0! geren_+ciamento,+te!
gerencial_0,-1! gerir_-2+encia,-2+ente! german_0,-6+alemanha! germanofil_+ia,+ico! gerontofil_+ia,+ico!
gestual_0,-3+o! gigante_-6+randeza,0! ginchar_-2+o,-1+dor! ginga_+da,+dor! girar_-2+o,-1+dor! glacia_0,-5+elo!
glamouros_-2,0! glandular_0,-1! gliss_+ada,+ador! globaliza_+cao,+dor! glosar_-1,-1+dor glut_+onaria,+ao!
glute_0,-5+nadegas! gnostic_+ismo,+o! gofr_+agem,+ador! goiv_+adura,+ador! golea_+da,+dor! golfistic_0,-5+e!
golpea_-1,+dor! goma_+gem,+dor! gonidial_0,-2+o! gord_+ura,+o! gordur_+a,+oso! gorje_+io,+ador!
gostos_+ura,+o! governamental_0,-7+o! gradual_+ismo,0! gramar_-1+do,-1+dor! gramatic_0,+a!
granar_!-1+cao,-1+dor! grand_+eza,+e! granje_+io,+ador! grao_!-1+ndeza,-1+nde! grasn_+ada,+ador!
gratif_+icacao,+icador! gratu_+idade,+ito! grava_+cao,+dor! gravav_-1+cao,0! grave_-1+idade,0! grego_0,-2+cia!
gremist_0,-2+o! greta_+dura,+dor! grevist_-3+e,+a! grifar_-2+o,-1+dor! grimp_+agem,+ador! grip_+e,+ado!
grita_-1+o,+dor! gross_+ura,+o! grotesco_!0,0! gruda_+dura,+dor! grunhi_+dela,+dor!
grup_?0?+amento,?+o?+ador! guapo_-1+eza,0! guard_+a,+ador! guerre_?0?-1+a,?-1+a?+iro! gui_+amento,+a!
guiar_-1,-1! guisar_-1+do,-1+dor! gular_0,-4+arganta! gutural_0,-6+arganta! habita_+cao,+nte!
halometr_+ia,+ico! haver_-5+existencia,0! hediond_+ez,+o! helenic_0,-2+os! helicoida_0,-9+helice!
heliotrop_+ia,+ico! hematic_0,-7+sangue! hemiedr_+ia,+ico! hemisferic_0,-1+o! hemostatic_-3+se,+o!
hepatic_0,-7+figado! herdar_-3+anca,-2+eiro! hered_-2+anca,0! hermeneutic_+a,0! hermetic_-1+smo,0!
heroi_+smo,0! heroic_-1+smo,+o! heterolog_+ia,+ico! heteroplas_+ia,+ico! heuristic_+a,+o! hexagin_+ia,0!
hibrid_+ismo,0! hidatic_0,-1+de! hidraulic_0,+a! hidric_0,-6+agua! hidrostatic_-3+sia,+o! hieroglif_0,+o!
higienic_-2+e,+o! hilari_-6+comicidade,0! hinaianistic_+a,+o! hiperagud_+eza,+o! hiperbolic_0,-2+e!
hipertextual_0,-2+o! hipertrofia_-0,+dor! hipico_0,-6+cavalo! hipnotiz_-1+smo,+ador! hipocondr_+ia,+iaco!
hipocratic_0,-2+es! hipocri_+sia,+ta! hipostatic_-2+se,+o! hipoteca_-0,+rio! hipotetic_0,-3+se!
hipotireoid_+ismo,+e! hissop_+ada,+ador! hister_+ia,+ico! histori_0,+a! historiografic_0,-1+a! histrion_+ice,0!
hodiern_-7+atualidade,+o! holoedr_+ia,+ico! homenagea_-1+m,+dor! homeostatic_-3+se,+o! homizi_+o,+ador!
homologo_-1+ia,0! homotal_+ia,+ico! homotip_+ia,+ico! hondurenh_0,-3+as! honr_+a,+ado! horr_+or,+ivel!

hospeda_+gem,+dor! hospitalar_0,-2! hospitale_-1+idade,0! humild_+ade,+e! humilha_+cao,+dor!
humorad_-2,+o! icar_!-1+mento,-1+dor! icosandr_+ia,0! idear_-2+ia,-1+dor! identic_-1+dade,+o! idilic_0,-1+o!
idiot_+ice,+a! idolatr_+ia,+ador! ignav_+ia,0! igneo_0,0! ignoran_+cia,+te! igualitar_+ismo,+io! ileso_0,0!
ilhar_-5+isolamento,-5+isolador! ilicit_+ude,+o! ilogic_0,+o! ilustra_+cao,+dor! imagina_+cao,+dor!
imaginari_-2+cao,0! imaginav_-1+cao,0! imagistic_+a,+o! imane_-1+idade,0! imanent_-3+encia,+e! imaterial_0,0!
imatur_+idade,+o! imediat_+idade,+o! imenso_-1+idao,0! imiscu_+icao,+idor! impacient_-1+cia,+e!
impacta_-1+o,+dor! impar_!+idade,0! imparissilab_+ismo,+o! imperar_-2+io,-1+dor! imperativ_-6+osicao,0!
imperi_+cia,+to! imperial_+ismo,0! imperman_+encia,+ente! impied_+ade,+oso! impingi_+dela,+dor!
impio_!-5+descrenca,0! importuno_-1+ice,0! imprens_+adura,+ador! imprevis_+ibilidade,+to!
imprevisiv_-1+bilidade,0! imprima_+dura,+dor! imprimi_-3+essao,-3+essor! improvisa_+cao,+dor!
impuro_-1+eza,0! imund_+ice,0! inadvertid_-2+encia,0! inativ_+idade,+o! inaudit_+ismo,+o! inaugural_-1+cao,0!
incauto_-7+imprudencia,0! incend_+io,+iario! incentiv_+o,+ador! incestuos_-3+o,0! incidental_-3+cia,0!
inclu_+sao,+sor! incolum_+idade,+e! incomoda_+cao,+dor! incongru_+encia,+ente! incorpore_+idade,+o!
increment_+o,+ador! inculc_+a,+ador! incult_+ura,+o! indaga_+cao,+dor! indebit_-7+improcedencia,+o!
indelica_+deza,0! indevid_0,0! indian_0,-1! indiferen_?+ca?+ciacao,+ciador! indigena_0,-4+o! indigenat_+o,+o!
indigest_+ao,0! indign_+idade,0! indiscre_-1+icao,+to! indumentari_0,+a! induvial_0,-1! inedit_+ismo,+o!
inepto_-2+cia,0! inercial_0,-1! inerr_+ancia,+ante! inert_-1+cia,+e! inescrupulos_-1,0! inexat_+idao,+o!
infama_-1+ia,+dor! infeccios_-3+ao,0! infecund_+idade,+o! infenso_0,0! infernal_0,-1+o! infinitiv_0,+o!
inflacionari_-6+ao,0! infle_+xao,0! inflig_-1+ccao,+idor! influenci_+a,+ador! informatic_0,+a!
infringi_-4+acao,-4+ator! ingenu_+idade,+o! inglori_0,+o! ingluvial_0,-2+o! ingrem_+idade,+e!
ingressa_-1+o,+dor! inguinal_0,-8+virilha! inicia_-1+o,+nte! inicial_0,-2+o! iniciativ_+a,+o! inimig_-1+zade,0!
injuri_+a,+oso! inobserv_+ancia,+ante! inocent_-1+cia,+ador! inocenta_-8+absolvicao,+dor! inoper_+ancia,+ante!
inospit_+alidade,+o! inquer_-2+iricao,-2+iridor! inquir_-2+erito,+idor! insalubr_+idade,+e! insano_-1+idade,0!
insatisf_+acao,+atorio! insensat_+ez,+o! inser_+cao,+idor! insignific_+ancia,+ante! insipi_+encia,+ente!
insobri_+edade,+o! insolv_+encia,+ente! insone_-1+ia,0! instanc_+ia,+iador! institucional_0,-6+icao!
instrui_-1+cao,-1+tor! insuave_-1+idade,0! insubordina_+cao,+do! insubstancia_0,0! insuet_+ude,+o!
insult_+o,+uoso! insurrecional_0,-6+icao! integro_-1+idade,-1+o! inteira_+cao,+dor! inteiro_-1+eza,0!
intelectiv_0,-2+o! intenso_-1+idade,0! inter_0,0! intercambi_+amento,+ador! intercorr_+encia,+ente!
interessa_-1+e,+nte! interesse_-0,+iro! intereste_0,-9+estrela! interin_+idade,+o! interligar_-1+cao,-1+dor!
intermedia_+cao,+rio! interno_0,-2+ior! interpela_+cao,+dor! interrog_+atorio,+ador! interromp_-3+upcao,+edor!
intersec_-1+cao,0! intersticial_0,-2+o! intertropical_0,-2+o! interv_+encao,+entor! intestinal_0,-2+o!
intimida_+cao,+dor! intoc_0,0! intolera_+ncia,+nte! intra_0,0! intransitiv_+idade,+o! intrig_+a,+ante!
introspec_+cao,+tivo! intru_+sao,+so! intui_+cao,+dor! intui_+cao,+tivo! inunda_+cao,+dor! inundav_-1+cao,0!
invad_-1+sao,-1+sor! invectiv_+a,+ador! invejar_-1,-2+oso! invejav_0,-1! invent_-1+cao,+or! inventaria_+do,+dor!
inverna_-1+o,+dor! invers_+ao,0! invict_-3+encibilidade,+o! ir_!-1+da,-1+do! iracund_+ia,0! irad_-1,0!
irar_!-1,-1+do! irial_0,-2+s! ironi_+a,+zador! irracional_0,0! irreal_+ismo,0! irrealis_+mo,+ta! irresol_+ucao,+vido!
irrita_+cao,+do! isent_-1+cao,+ador! islamic_-1+smo,+o! isobatic_0,-2+a! isolar_-1+mento,-1+dor!
isostatic_-3+sia,+o! isotrop_+ia,+ico! italian_0,-1! itinera_-7+caminho,+nte! janist_-1+mo,+a! janot_+ice,+a!
jantar_-1,-1+dor! jasminace_0,-4+m! jaze_-1+imento,+nte! jazzistic_0,-5! jejunal_0,-2+o! jeremia_-7+choro,+dor!
jogar_-2+o,-1+dor! jovem_-4+uventude,-1+m! jucund_+idade,+o! judicia_0,-5+stica! judicial_0,-1+rio!
jugal_0,-2+o! jugular_0,-7+garganta! jumental_0,-2+o! juntar_-6+proximidade,-1+dor! junto_-5+proximidade,0!
jurament_+o,+ador! jurid_0,-3+stica! jurisperi_+cia,+to! justa_-1+ica,0! justic_+a,+eiro! justificav_-1+cao,0!
justo_-1+eza,0! labial_0,-2+o! labil_+idade,0! labirinti_0,-1+o! lacar_!-2+o,-1+dor! laconic_-1+smo,+o!
lacrimal_0,-6+grima! lacteo_0,-5+eite! ladin_+ice,+o! ladrao_-6+roubo,0! ladrar_-6+latido,-1+dor! lagunar_0,-1!
laico_0,0! laje_+amento,+ador! lambar_+ice,+eiro! lambe_-1+ida,+dor! lambisc_+ada,+ador! lambisgo_+ice,+oia!
lambris_+amento,+ador! lambuz_+ada,+ador! lamech_+ice,+a! lanar_0,-3! lanchar_-2+e,-1+dor! larga_+da,+dor!
largo_-1+ura,0! laringe_0,-1+e! lasciv_+ia,0! lasso_-1+idao,0! lastima_-0,+dor! lastimav_0,-1! latir_-1+do,-1+dor!
lato_-2+rgura,0! laure_?0?+a,?-4+ouro?+ador! lava_+gem,+dor! lavav_-1+gem,0! lavra_-0,+dor!
leciona_-6+icao,+dor! ledo_-4+alegria,0! legenda_+gem,+dor! legendari_0,-2! leigo_-1+uice,0! leiloa_-2+ao,+dor!
leiteir_0,-2! lembra_+nca,+dor! lendari_-2,+o! lento_-1+idao,0! leonistic_-7+ao,0! lepros_-2+a,0! leptorrin_+ia,0!
ler_!-1+itura,-1+itor! lerdo_-1+eza,0! lesa_+o,+dor! lesar_-2+o,-1+dor! lesbic_-1+anismo,+o!
leseir_+a,-6+preguicoso! leto_!0,+nia! letrado_-7+erudicao,0! leucocitari_0,-3+o! levar_!-5+transporte,-1+dor!
leve_+za,0! levian_+dade,+o! lexical_0,-2+o! libertari_0,-2+cao! liberto_-2+dade,0! librar_-6+equilibrio,-1+dor!
licito_-1+ude,0! lidar_!-1,-1+dor! lider_+anca,0! ligar_-1+cao,-1+dor! ligeir_+eza,+o! ligular_0,-1! lima_+da,+dor!
limenh_0,-3+a! liminar_0,-3+ar! liminar_-7+antecedencia,0! limp_+eza,+ador! linace_0,-3+ho! lindo_-5+beleza,0!
linfatic_0,-1+a! lingual_0,-1! liquida_+cao,+dor! lisboeta_0,-3+a! liso_-1+ura,0! lisonjea_-1+io,+dor!
lista_+gem,+dor! listra_+do,+dor! litagog_+ia,0! literari_+edade,+o! litig_+io,+ante! litora_0,0!
livra_-3+berdade,-3+bertador! livre_-3+berdade,0! livresc_0,-3+o! lobreg_+uidao,+o! local_0,-0! locatici_-4+cao,0!
logic_0,+a! logistic_+a,+o! logo_!-4+precisao,0! lombar_0,0! londrin_0,-2+es! longe_!-5+distanciamento,0!
longev_+idade,+o! longinq_-7+distanciamento,0! longitudinal_0,0! longo_-1+uidao,0! lorp_+ice,+a! loteric_0,-1+a!
louca_+nia,+ao! louco_-1+ura,0! lour_+idao,-4+alourador! lucra_-1+o,+dor! lucrativ_+idade,+o!
ludibria_-1+o,+dor! ludic_-5+divertimento,+o! lugubr_+idade,+e! lunar_0,-3+a! lusitan_0,-6+isboa! lustral_0,0!
luta_-0,+dor! luxuos_-3+o,0! luz_-3+brilho,+ente! macabr_+ismo,+o! macac_-1+quice,+o! macacal_0,-2+o!
macambuz_+ismo,+io! macanj_+ice,+o! macaquea_-2+ice,+dor! macho_0,0! machuca_+do,+dor! macic_+ez,+o!
macio_-1+ez,0! maconic_0,-2+aria! macrofil_+ia,+ico! macular_-1,-1+dor! macunaimi_0,-1+a! madrac_+ice,+o!
madur_-3+turidade,-5+amadurecedor! magistral_0,-8+estre! magnetostatic_-3+sia,+o! magnific_+encia,+o!
magno_-1+itude,0! mago_!-1+ia,0! magoa_-0,+dor! magro_-1+eza,0! maior_+ia,0! majest_+ade,+oso!
majoritari_0,-8+ioria! mal_!-3+erro,0! malaguenh_0,-4+a! malandr_+agem,+o! maldit_-1+cao,+o!
maldos_-2+ade,0! malear_0,-4+rtelo! malefic_-1+cio,+o! malhar_-1+cao,-1+dor! malquerenc_+a,+ador!
malson_+ancia,+ante! maltrat_+o,+ador! maluc_-1+quice,+o! malvace_0,-2! malvad_+ez,+o! mamifer_0,+o!
manar_!-1+cao,-1+dor! mancha_-0,+dor! manda_-1+o,+nte! mandibular_0,-1! mandr_+ia,+iao! mandran_+ice,+a!
maneira_+da,-1+o! manifest_+acao,+ador! maninh_+ez,+o! manobra_-0,+dor! manso_-1+idao,0!
mante_-2+utencao,+dor! mantric_0,-2+a! manual_0,0! manufatu_+ra,+rador! manuse_+io,+ador!
maquiar_-1+gem,-1+dor! maquiaveli_+smo,+co! maravilhos_-2+a,0! march_+a,+ador! marcial_0,-7+luta!
marcian_0,-4+te! margea_+cao,+dor! marin_?0?+acao,?-2?+ador! marinar_-1+da,-1+dor! marinha_+gem,0!
marinhe_0,-4! marinho_0,-4! marit_0,-2! marital_0,-5+trimonio! marmor_0,+e! marmorei_-1+acao,-1+ador!

marquesal_0,-2! marralh_+ice,+eiro! marroquin_0,-4+cos! masculin_+idade,+o! masculo_-1+inidade,0!
massacr_+e,+ador! massiv_+idade,+o! matar_-5+assassinato,-1+dor! matematic_0,+a! material_0,0!
matraquea_-4+ca,+dor! matreir_+ice,+o! matrici_0,-2+z! matricul_+a,+ador! matriz_-1+z,+ador! mau_-1+ldade,0!
mavort_+ismo,+ico! maxilar_0,-1! maxim_0,+o! mazomb_+ice,+eiro! mear_!-1+cao,-1+dor! mecanicis_+mo,+ta!
medei_-2+iacao,-2+iador! median_+ia,0! mediastinal_0,-2+o! medic_?0?+amento,?+ina?+o! medicin_+a,-2+o!
medieval_0,0! mediocr_+idade,+e! mediterr_0,aneo! mediu_+nidade,+m! medonho_-7+feiura,0!
medra_+nca,+dor! medros_-3+o,+o! medular_0,-1! megaloman_+ia,+iaco! meigu_+ice,+o! meio_!-2+tade,0!
melancol_+ia,+ico! melancoli_+a,+co! melani_+a,+co! melhor_!-6+superioridade,0! melhora_-0,+dor!
melhore_-1+a,-1+ador! meliflu_+idade,+o! melindr_+e,+oso! melodios_-2+a,0! memorav_0,-2+ia!
mendig_-1+cancia,+o! mene_+io,+ador! menor_+idade,0! menoscab_+o,+ador! menosprez_+o,+ador!
mensal_0,-4+s! mental_?0?+izacao,-2+e! mentir_+a,+oso! mercanti_0,-8+comercio! mercar_-1+do,-1+dor!
merend_+a,+eiro! meridional_0,0! merito_0,0! mesclar_-1+gem,-1+dor! mesmo_-1+ice,0! mesofalangeal_0,-2!
mesorrin_+ia,0! mesquinh_+ez,+o! mestic_+agem,+o! mestr_+ia,+e! mesur_+ice,0! metaboli_+smo,+zador!
metaforic_0,-2+a! metalic_0,-2! metastatic_-3+se,+o! meter_-2+icao,-2+idor! metodi_0,-1+o! metrico_0,-3+o!
metropol_0,+e! mexe_+cao,+dor! mexeric_+o,-1+queiro! mexican_0,-2+o! miar_-1+do,-1+dor! microfil_+ia,0!
micross_+ismo,ismico! milagros_-2+e,0! milenar_!0,0! milita_+ncia,+nte! mimalh_+ice,+o! mimar_-2+o,-1+dor!
mimetiz_-1+smo,+ador! mimic_+a,+o! minar_!-5+abalo,-1+dor! mineira_+cao,-1+o! mineral_0,-2+io!
minerval_0,-1! minguar_-1,-1+dor! minimal_0,0! ministerial_0,-2+o! minora_+cao,+dor! minorit_0,-1+a!
minoritari_0,-4+a! minutar_-1,-1+dor! mirabol_+ancia,+ante! miracul_-5+lagre,0! mirar_-1,-1+dor!
miserav_-2+ia,0! misericordios_-2+a,0! mistic_+ismo,+o! mistur_+a,+ador! mitico_!0,-3+o! mitocondria_0,-1+o!
miudo_-1+eza,0! mixuruc_-1quice,+a! mobilar_-2+ia,-2+iador! mobiliar_!-1,-1+dor! modela_+gem,+dor!
moderant_+ismo,+e! modesto_-1+ia,0! mofar_-2+o,-1+dor! molar_!0,0! molda_+gem,+dor! moldur_+agem,+ador!
mole_!+za,0! moleca_+gem,-2+que! molecular_0,-1! moleng_-2+za,+o! molha_+cao,+dor! monastic_+ismo,+o!
monda_-0,+dor! monetari_0,-6+eda! mongol_0,+ia! mono_!0,0! monogam_+ia,+ico! monolitic_0,0!
monolog_+o,+ador! monopod_+ia,+e! monstr_+uosidade,+o! montanh_0,+a! morador_-3+dia,0!
morar_-1+dia,-1+dor! morcegal_0,-2+o! morde_-1+ida,+dor! mordisca_+cao,+dor! morfossintatic_-3+xe,+o!
morgar_-1+da,-1+dor! moribun_-7+agonia,0! morn_+idao,-4+amornador! moroso_-1+idade,0! morre_-2+te,-2+to!
mortif_0,-2+e! morto_-1+e,0! mortuari_0,-4+e! mostr_+a,+ador! mostra_-0,+dor! mover_-2+imento,-1+dor!
muculman_+ismo,+o! muda_+nca,+dor! mudo_-1+ez,0! mulat_+ice,+o! multa_-1+a,+dor! multi_0,0!
multiplice_-1+idade,0! mundial_0,-3+o! municipal_0,-2+io! munific_+encia,+ente! mural_0,0!
murcha_+mento,+dor! muscular_0,-2+o! music_+a,+o! mutuo_-5+reciprocidade,0! nadar_-3+tacao,-1+dor!
nambiquar_0,+as! namora_-1+o,+dor! narcis_+ismo,0! narcisa_+mento,+dor! narra_+cao,+dor! nasal_0,-3+riz!
natalin_0,-2! nativis_+mo,+ta! naturaliz_+acao,+ador! naufrag_+io,+o! nause_+a,+ante! naval_0,-2+io!
navegar_-1+cao,-1+dor! navicular_0,-1! necessit_-1+dade,+ado! nedio_!-1+ez,0! negacea_-2+a,+dor!
negligen_+cia,+te! negocial_0,-2+o! negrei_0,-2+o! negro_-1+idao,0! neofil_+ia,0! nervos_+ismo,+o! neural_0,0!
neurotic_0,-3+se! neutr_+alidade,+o! neva_-1+e,-1+e! nicar_!-1+da,-1+dor! nimi_+edade,+o!
ninar_!-5+acalento,-1+dor! niponic_0,-7+japao! niquen_-2+ice,0! nirvanic_0,-2+a! nitid_+ez,0! nitric_-6+acidez,+o!
niveo_0,-4+eve! nobre_+za,0! noctivag_0,-6+ite! noiva_+do,-1+o! nojent_-3+o,+o! nomea_+cao,+dor!
nomenclatori_0,-3+ura! nominal_0,0! nordest_0,+e! nordeste_!0,0! noroeste_!0,0! norte_!0,0! nortenh_0,-2!
noticia_+rio,+dor! noto_+riedade,+rio! noturn_0,-4+ite! novel_-2+idade,0! noveles_0,-2+a! novelistic_0,-5+a!
novidadeir_+ice,+o! novilunar_0,-2+io! novo_-1+idade,0! nu_!+dez,0! nublar_-4+vem,-1+dor! nucal_0,-1!
nucular_0,-6+oz! nulo_-1+idade,0! numeros_-7+abundancia,0! nupcia_0,+s! nutr_+icao,+idor! obceca_+cao,+dor!
obeso_-1+idade,0! objetal_0,-2+o! obscurant_+ismo,+e! obscuro_-1+idade,0! obseca_+cao,+dor!
obsequios_-1,0! observav_-1+cao,0! obsolet_+ismo,+o! obstar_-1+ncia,-1+nte! obstru_+cao,+idor!
obte_+ncao,+dor! obtura_+cao,+dor! obtus_+idade,+o! obvi_+edade,+o! ocasional_-9+eventualidade,0!
ocasionar_-9+efeito,-1+dor! occipital_0,-3+cio! oceanic_0,-2+o! ocorre_+ncia,+nte! ocular_0,-5+lho!
ocult_+acao,+ador! odiar_!-2+o,-1+dor! odios_-1,0! ofega_-1+o,+nte! ofende_-2+sa,-2+sivo! ofensiv_0,-2+a!
oficial_!0,0! ofidi_0,+o! olhar_-1+da,-1+dor! olimpic_0,-1+adas! olvid_+o,+ador! omin_-4+agouro,-4+agourento!
ondee_-1+amento,-1+ador! ondei_-1+amento,-1+ador! ondula_+cao,+nte! onivor_+idade,+o!
onomatic_-8+nome,+o! opac_+idade,+o! operistic_0,-5+a! opinar_-2+iao,-1+dor! opiniatic_0,-3+o!
opta_-2+cao,+nte! optic_+idade,+o! oracula_-3+ao,-1+o! orar_!-1+cao,-1+dor! orbitari_0,-2! ordenh_+a,+ador!
orfa_+ndade,+o! organolog_+ia,+ico! orgulh_+o,+oso! origi_-1+em,+nador! original_+idade,0! oriund_-3+gem,0!
orlar_-1+dura,-1+dor! ornar_-1+mento,-1+dor! orquestra_+cao,+dor! orquestral_0,-1! ortomolecular_0,-1!
osmolar_0,-2! otario_0,0! otico_!0,-1+a! otimo_0,0! otoman_0,-6+turquia! ousa_+dia,+do! outonal_0,-2+o!
outorg_+a,+ador! ouv_-3+audicao,+inte! ouvi_-4+audicao,+nte! oval_!0,0! ovin_0,-2+elha! pacat_+ez,+o!
pachol_+ice,+a! pacif_?0?+icacao,?-3+z?+icador! pair_+o,+ador! palacian_+ismo,+o! palatal_0,-2+o!
palerm_+ice,+a! palestin_0,+a! palestr_+a,+ante! palmar_0,-1! palpebral_0,-2! palr_+ice,+ador! palurd_+ice,+io!
panamenh_0,-3+a! pancreatic_0,-1+a! panflet_+agem,+ista! panic_0,+o! panoramic_0,-2+a! pantanei_0,-2+al!
papal_0,-1! papalv_+ice,+o! papeat_+a,+eiro! paquera_-0,+dor! par_!+idade,0! parabol_0,+e!
paradisiaco_0,-7+iso! paradoxal_0,-2+o! parafin_+agem,+ador! paralel_+ismo,+o! paralitic_-3+sia,0!
paranaens_0,-3! parangon_+agem,+ador! paranoic_-1+a,+o! parar_-1+da,-1+dor! parasit_+ismo,+a!
parasitari_0,-2! parco_-1+imonia,0! parece_-6+aparencia,-1+ido! paregor_+ia,+ico! parent_+esco,+e!
pari_-1+to,+dor! paritari_0,+o! parlamenta_0,-1+o! parodi_+a,+ista! parol_+ice,+ador! paronim_+ia,+ico!
paroqui_+amento,+ador! paroquial_0,-1! parox_+ismo,+al! participia_0,-1+o! partilha_-0,+dor!
partir_-1+da,-1+dor! parvo_+ice,0! pascac_+ice,+io! pascoal_0,-1! pasmar_-1+ceira,-1+do!
passa_?0?+gem,+geiro! passad_0,+o! passional_0,-7+ixao! passivo_-1+idade,0! pastorea_-1+io,-1+iro!
pastoril_0,-2! pateg_+uice,+o! patelar_0,-1! patente_!-7+evidencia,0! patetic_+ismo,+o! patina_+cao,+dor!
patinha_+gem,+dor! patriarca_0,-1+a! patristic_+a,+o! patronal_0,-4+ao! patrulha_+mento,+dor!
paulist_0,-7+sao=paulo! pausa_-0,+dor! pautar_-1,-1+dor! pecar_!-1+do,-1+dor! pechincha_-0,+dor!
pecuniari_0,-2! pedala_+gem,+dor! pedant_+ismo,+e! pedi_-2+ticao,+nte! pedir_-1+do,-1+nte! pedregos_-4+a,0!
pegar_-1+da,-1+dor! pegural_0,-2+eiro! peitar_-1+da,-1+dor! pelar_-1+dura,-1+dor! pele_+dura,+dor!
pelech_+o,+ador! pelintr_+ice,+a! pelud_-2+o,0! penar_!-1,-1+do! pendur_+a,+ador! peninsular_0,-1!
penitenciari_0,+a! penso_!0,0! pentecostal_0,-2+e! pequen_+ez,+o! peralt_+ice,+a! perceb_-1+pcao,+edor!
percent_+agem,0! percus_+sao,+sor! percut_-1+ssao,-1+ssor! perde_-1+a,+dor! perdoa_-2+ao,+dor!
peremp_+cao,0! peremptor_+iedade,0! perfaze_-7+completamento,+dor! perfid_+ia,+ico! performat_-1+nce,0!
perfum_+e,+ador! pergunta_-0,+dor! perici_+a,-2+to! periferic_0,-1+a! perigar_-2+o,0! perigos_-1,0!

perineal_0,-2+o! periodic_+idade,+o! periosteal_0,-2+o! perit_-1+cia,+o! peritonial_0,-2+o! perjur_+io,+ador!
perlux_+idade,+o! permane_+ncia,+nte! permea_-1+io,+dor! permissiv_+idade,+o! pernambuc_0,+o!
pernostic_-9+afetacao,+o! peroneal_0,-2+o! perquir_+icao,+idor! perreng_+agem,+ue! perseg_+uicao,+uidor!
persever_+anca,+ador! persic_0,-1+a! perspectiv_+a,+o! pertenc_-3+inencia,+ente! pertina_+cia,0!
perto_-5+proximidade,0! pervers_+idade,0! pesa_+gem,+dor! pesaro_-1,0! pesca_+ria,+dor! pesg_+a,+ador!
pesquei_0,-4+ca! pesquis_+a,+ador! pessoal_0,-1! petar_!-5+mentira,-5+mentiroso! petisc_-1+quice,+o!
petre_0,-3+dra! petroli_0,-1+eo! petul_+ancia,+ante! pianistic_0,-5+o! picant_0,0! pichar_-1+cao,-1+dor!
picot_+agem,+ador! pictoric_0,-6+ntura! piegu_+ice,+as! pifar_-1+da,-1+do! pilh_+agem,+ador!
pilhar_-1+gem,-1+dor! pilota_+gem,-1+o! pingar_-2+o,-1+nte! pinic_+ada,+ador! pinta_-1+ura,-1+or!
pior_!-4+inferioridade,0! piora_-0,+dor! piquet_+agem,+ador! pirata_+ria,0! piratic_0,-2+o! pirotecni_+a,+co!
pirron_+ice,+ico! pisa_+da,+dor! piscar_-1+da,-1+dor! pisceo_0,-5+eixe! piscos_0,-5+eixe! pisoa_+mento,+dor!
pitoresco_!0,0! placentari_0,-2! plagi_+o,+ador! plagia_-1+o,+dor! planalt_0,+o! planar_!-1+da,-1+dor!
planch_+ada,+ador! planea_-1+jamento,-1+jador! planeja_+mento,+dor! planetari_0,-2! plang_+encia,+ente!
planisferic_0,-1+o! platicefal_+ia,0! platina_+gem,+dor! platirrin_+ia,0! platonic_0,+o! plebe_+idade,+u!
pleitea_-2+o,+dor! pleno_-1+itude,0! pleural_0,-1! plota_+gem,+dor! plumb_+ismo,+ico! plumbe_0,-6+chumbo!
pluri_0,0! pluvial_0,-7+chuva! pneumatic_0,-9+ar! pobre_+za,0! poda_-0,+dor! poder_-0,0! podr_+idao,+e!
poeirent_-3+a,+o! poetic_0,-3+sia! polemi_+ca,+co! policial_0,-1! poligastric_+idade,+o! poligin_+ia,0!
poligonal_0,-2+o! polimatic_0,-1+a! polimetr_+ia,+ico! polir_!-1+mento,-1+dor! politic_0,a! poltrona_+ria,+ao!
polu_+icao,+idor! ponca_+gem,+dor! ponte_+io,+ador! pontific_0,+ie! populos_-2+acao,0!
por_!-1+sicao,-3+colocador! porcent_-7+percentagem,0! porco_-1+aria,0! porno_!+grafia,0! poroso_-1+idade,0!
portar_-2+e,-1+dor! portenh_0,-7+buenos-aires! portuari_0,-4+o! posar_-1+da,-1+dor! positivo_-1+idade,0!
possesiv_-3,+o! possu_-1+e,+idor! posta_+gem,+dor! postal_0,-6+correio! postico_0,0! potencial_+idade,0!
pouco_-5+escassez,0! poupa_+nca,+dor! pous_+ada,+ador! povoa_+cao,+dor! pragmatic_+a,0! praiei_0,-3!
pratea_+cao,+dor! pratica_-0,+nte! pratico_-1+idade,0! prazeros_-2,0! prebostal_0,-2+e! precari_+edade,+o!
precave_-7+cautela,-1+ido! preceptor_+ia,0! precisa_-7+necessidade,-7+necessitado! precoc_+idade,+e!
preconceit_+o,0! predial_0,-2+o! predispo_+sicao,+sto! predomin_+io,+ante! preemp_+cao,+tivo!
prefacia_-1+o,+dor! prega_+gem,+dor! pregn_+ancia,+ante! prego_+amento,+ador! preguicos_-2+a,0!
prejudic_-3+izo,+ador! preliminar_-7+cedencia,0! prematur_+idade,+o! premer_-3+ssao,-1+nte!
premiar_-2+o,-1+dor! premir_-3+ssao,-2+ente! premoni_+cao,+torio! premun_+icao,+itor! prende_-4+isao,+dor!
prenhe_+z,0! preponder_+ancia,+ante! prepostero_-1+idade,0! prepucial_0,-2+o! prerrog_+ativa,+ativo!
presbiteral_0,-2+o! prescindi_-9+dispensa,-9+dispensador! presen_+ca,+te! presentea_-1,+dor!
presidencial_0,-4+te! pressagi_+o,+oso! prestigi_+o,+ador! presum_-1+ncao,-1+ncoso! presuncos_-2+ao,0!
pretend_-1+sao,+ente! preter_+icao,+idor! pretir_-2+ericao,-2+eridor! preto_-1+idao,0! pretori_0,-1!
previn_-2+encao,-2+enidor! preza_-5+estima,+dor! primacial_0,-4+z! primar_-1+zia,-2+oroso! primari_+edade,+o!
primitiv_+ismo,+o! primo_!0,0! primogenit_+ura,+or! principal_0,0! principi_0,+o! prioral_0,-2! privad_0,0!
privilegia_-1+o,+dor! probatori_-6+va,+o! probo_-1+idade,0! procede_-1+imento,-2+ssador! proconsular_0,-2!
procura_-0,+dor! profano_-1+acao,0! profer_+icao,+idor! profesa_-3+issao,-1+sor! professoral_0,-2!
profet_-1+cia,+a! profir_-2+eridor,-2+eridor! profunda_-8+aprofundamento,+dor! profundo_-1+idade,0!
prognatic_0,-1+a! prognostic_+o,+ador! programav_-1+cao,0! progred_-1+sso,-1+ssivo! progressiv_+idade,+o!
proletari_0,+o! prolif_+eracao,+ico! prometeic_0,-2+u! prometer_-3+ssa,-1+dor! promulga_+cao,+nte!
pronotal_0,-2+o! pronto_-1+idao,0! pronu_+ncia,+nciante! propaga_+cao,+dor! propicia_+cao,+dor!
propositad_0,-2+o! propul_+sao,+sor! prosa_?0?-0,?-0?+dor! prosis_+mo,+ta! prosper_+idade,+o!
prostatic_0,-1+a! proteic_0,-1+na! protela_+cao,+dor! protend_-1+sao,-1+sor! proterv_+ia,0! protesta_-1+o,+nte!
protetoral_0,-2+do! protocolar_-2+o,-1+dor! protopatic_+a,+o! provar_-1,-1+dor! provencal_0,-1! proverbia,0,0!
providencia_?0?-0,?-0?+dor! provincial_0,-1! provisori_0,+o! proxenet_+ismo,+a! proxim_+idade,+o!
psicotic_0,-3+se! psiqu_0,+e! psiquiatri_0,+a! ptotalitar_+ismo,+io! puber_+dade,+e!
public_?0?+acao,?-5+ovo?+ador! publicit_0,-1+dade! pujan_+ca,+te! pular_!-2+o,-1+dor! pulcr_+itude,+o!
pulmonar_0,-4+ao! pulpar_0,-5+olpa! pulsa_+cao,+dor! pung_-1+cao,-1+idor! puni_+cao,+dor! puritan_+ismo,0!
puro_-1+eza,0! puxa_+da,+dor! quadrangular_0,0! quadratic_0,-3+do! quadricula_+do,+dor! quant_+idade,0!
quebr_+a,+ador! queda_-5+permanencia,+dor! queima_+da,+dor! queix_+a,+oso! queixal_0,-2+o!
quente_!-6+calor,0! querel_+a,+ador! querer_!-6+desejo,-1+nte! querosen_+agem,+ador!
questi_+onamento,?0?+onador! quicar_-3+que,-1+nte! quilometra_+gem,+dor! quiropratic_+a,+o!
quita_+cao,+dor! quitenh_0,-3+o! rabea_+da,+dor! rabec_+ada,+ador! rabic_0,-1+a! rabin_+ice,+o!
rabisca_-1+o,+dor! rabug_+ice,+ento! racha_+dura,+dor! racia_0,-2+a! raciocin_+io,+ador! racional_0,-6+zao!
racist_-1+mo,+a! radial_0,-4+io! radiativ_+idade,+o! radicalar_0,-2! radicular_0,-1! radioativ_+idade,+o!
raiar_!-0,0! raiva_-0,-1+oso! raivos_-2+a,0! ralass_+aria,+o! ralha_-0,+dor! ramalh_+ada,+ador! ramular_0,-4+o!
rang_+ido,+edor! ranhet_+ice,+o! ranzinz_+ice,+a! rapa_+da,+dor! rapid_+ez,0! rapina_+gem,+dor!
rapta_-1+o,+dor! raquiti_+smo,0! raro_!-1+idade,0! rasar_-1+dura,-1+nte! rasg_+amento,+ador!
raso_!-4+profundidade,0! raspa_+gem,+dor! rastej_+o,+ador! ratea_-1+io,+dor! readotar_-3+cao,-1+dor!
readquir_-5+quisicao,+idor! reajusta_-1+e,+dor! realca_-1+e,+dor! realizav_-1+cao,0! reaparec_-2+icao,+edor!
reassum_-1+ncao,+idor! reave_-3+cuperacao,-3+cuperador! rebate_-1+ida,+dor! rebatiza_-2+smo,+dor!
rebel_+iao,+de! rebeld_+ia,+e! reboc_-1+que,+ador! rebola_+do,+dor! recalca_-2+que,+dor! recalcula_-1+o,+dor!
recambi_+o,+ador! recauchut_+agem,+ador! receita_-0,+nte! recem_0,0! recent_+idade,+e! receos_-2+io,0!
recessiv_+idade,+o! rechea_-1+io,+dor! rechin_+o,+ador! recicla_+gem,+dor! reciclav_-1+gem,0!
reciproc_+idade,+o! recobri_+mento,+dor! recomenda_+cao,+dor! recomendav_-1+cao,0! recompens_+a,+ador!
reconfort_+o,+ador! reconsert_+o,+ador! reconstru_+cao,+tor! recont_+agem,+ador! recopia_-0,+dor!
record_+e,+ista! recorda_+cao,+dor! recrut_+amento,+ador! recua_-1+o,+dor! recubr_-3+obrimento,-3+obridor!
recusa_-0,+dor! redarg_+uicao,+uidor! redesenha_-1+o,-1+ista! redig_-2+acao,-2+ator! redim_-2+encao,-2+entor!
rediscut_-1+ssao,+idor! redond_+eza,-6+arredondador! redr_+a,+ador! redund_+ancia,+ante!
reeleg_-1+icao,-1+itor! reemend_+a,+ador! reenlac_+e,+cador! reestamp_+a,+ador! reestuda_-1+o,+nte!
reexam_+e,+inador! reextrad_+icao,+itor! refals_+amento,+ador! refaze_-1+imento,+dor! refeit_-1+cao,+o!
referencia_?0?-0,?-0?+dor! referend_+a,+ador! refilma_+gem,+dor! reflet_-1+xao,+idor! reflexi_-1+ao,+ionador!
reflux_+a,-1+ente! refoga_+do,+dor! reforc_+o,+ador! reform_+a,+ador! refra_+cao,+tador! refratari_0,+o!
refresca_+mento,+dor! refuga_-1+o,+dor! refugia_-1+o,+do! refundi_+cao,+dor! regar_-1,-1+dor!
regatar_-1+gem,-1+dor! rege_+ncia,+nte! regel_+o,+ador! regiona_0,-3+ao! registr_+o,+ador!
regrava_+cao,+dor! regred_-1+ssao,-1+ssador! regres_+so,+sador! regressa_-1+o,+dor! regulav_-1+cao,0!

reina_-1+o,-2! reinicia_-1+o,+dor! reinquir_+icao,+idor! reins_+ercao,+eridor! reinven_+cao,+tor! reitoral_0,-2!
rejeitar_-3+cao,-1+dor! relac_+ao,+ionador! relancea_-1,+dor! relata_-1+o,-1+or! relativ_+idade,0!
relembra_+nca,+dor! reler_!-5+leitura,-5+leitor! religios_+idade,0! relut_+ancia,+ante! reluz_-5+brilho,+ente!
remanisc_+o,+ador! remar_-1+da,-1+dor! remat_+e,+ador! remedia_-1+o,+dor! remend_+agem,+ador!
remene_+io,+ador! remess_+a,+ador! remete_-2+ssa,+nte! remit_-1+ssao,+idor! remoa_+gem,+dor!
remocante_-7+juvenescimento,-7+juvenescedor! remodela_+gem,+dor! remoinh_+o,+ador! remolh_+o,+ador!
remond_+agem,+ador! remurmuri_+o,+ador! renal_0,-4+im! rende_-1+imento,0! rendos_-2+a,0! reng_+ueira,+o!
renomad_0,-2+e! renovav_-1+cao,0! renu_+ncia,+nciante! renuncia_-0,+nte! repass_+e,+ador!
repercut_-1+ssao,+idor! repergunta_-0,+dor! repet_+icao,+idor! repic_-1+que,+ador! repint_+e,+ador!
repis_+a,+ador! repleto_0,0! repous_+o,+ador! repreend_-1+sao,+edor! repreg_+o,+ador!
representativ_+idade,+o! repris_+e,+ador! reproduziv_-3+cao,0! rept_+o,+ador! republic_0,+a! repudi_+o,+ador!
repugn_+ancia,+ante! repuls_+a,+ador! repux_+o,+ador! requebr_+o,+ador! requint_+e,+ador!
requisi_+cao,+tante! rescald_+amento,+ador! resenh_+a,+ador! resenha_-0,+dor! reserv_+a,+ador!
resfoleg_+o,+ador! resgat_+e,+ador! resguard_+o,+ador! residual_0,-2+o! resin_+agem,+ador!
reslumbr_+ancia,+ante! resmunga_-1+o,+dor! resolut_-1+cao,0! respald_+o,+ador! respanc_+adura,+ador!
respectiv_0,+o! resping_+o,+ador! resplandec_+encia,+ente! responde_-3+sta,+dor! respost_+a,-2+ndente!
ressabia_-1+o,+dor! ressalg_+ada,+ador! ressalt_+o,+ador! ressalv_+a,+ador! ressemea_+dura,+dor!
ressoa_-1+o,+dor! resson_+o,+ador! ressorv_-1+cao,+edor! ressumbr_+o,+ador! ressur_+reicao,+gidor!
resta_-1+o,+nte! restaura_+cao,+dor! restri_+cao,+tor! result_+ado,0! resum_+o,0! retal_0,-2+o!
retanch_+oa,+ador! reticul_+agem,+ador! retilin_-3+dao,0! retini_+r,+dor! retir_+ada,+ador! retirar_-1+da,-1+dor!
retirav_-1+da,0! reto_!-1+idao,0! retom_+ada,+ador! retorce_-2+sao,+dor! retoric_+a,+o! retov_+amento,+ador!
retranc_+a,-1+queiro! retrocarg_+a,+ador! retroce_+sso,+dente! retruc_-1+que,+ador! retumb_+ancia,+ador!
retumb_+o,+ador! reumatic_-1+smo,+o! reunir_-1+ao,-1+dor! revela_+cao,+dor! revid_+e,+ador!
revira_+volta,+dor! revoa_+da,+dor! revolt_+a,+oso! revolucion_0,-3+ao! revolute_+io,+ador! reza_-0,+dor!
rezing_+a,+ador! ribomb_+ancia,+ante! rico_!-2+queza,-1+o! ridicul_+izacao,+o! rijo_-1+eza,0! rilha_+dura,+dor!
rimar_-1,-1+dor! ripa_+gem,+dor! rir_-1+so,-1+sonho! risc_+a,+ador! risonh_-2,+o! risonho_-3,0! ritmic_0,-2+o!
rixar_-1,-2+ento! rizoidal_0,-2+e! robotic_0,+a! robust_+ez,+o! roca_+dura,+dor! rochos_-2+a,0!
roci_+ada,+ador! roda_-1+ura,+dor! rodea_-1+io,+dor! rodopi_+o,+ador! roer_!-1+dura,-1+dor!
rogar_-2+o,-1+dor! rojar_-2+o,-1+dor! romancea_-1+,-2+ista! romances_0,-1! romano_0,-2! romeno_0,-1+ia!
ronca_-1+o,+dor! ronda_-0,0! rosace_0,-2! rosar_-1+do,-1+do! rosear_-6+ruborizacao,-1+ador! roseo_0,-2+a!
rosn_+ada,+ador! rotari_0,-1+y! rotativ_+idade,+o! rotinei_0,-2+a! rotul_+agem,+ador! rotund_+idade,+o!
roub_+o,-4+ladrao! rouco_-2+quidao,0! roxo_-1+idao,-4+arroxador! rubrica_-0,+dor! rude_!+za,0!
rudiment_+o,+ar! rufar_!-5+arrufo,-1+dor! ruflar_-2+o,-1+dor! ruidos_-1,0! ruim_-1+ndade,-1+m! ruinos_-2+a,0!
ruir_!-1+na,-4+arruinador! ruivo_!-1+idao,0! rumar_-2+o,0! rural_0,-5+campo! sabatic_0,-3+do!
saber_-5+conhecimento,-1+dor! sabid_-5+conhecimento,0! sabio_-2+edoria,0! saborea_+da,+dor!
sabot_+agem,+ador! sabuj_+ice,+o! sacan_+agem,+a! sacar_!-3+que,-1+dor! sacha_+dura,+dor!
saciar_-1+cao,-1+dor! sacrific_+io,+ador! sacud_+ida,+idor! sadi_+smo,+co! sadio_-3+ude,0! safad_+eza,+o!
safar_-5+desembaracamento,-1+do! sage_+z,_1+e! sai_+da,+dor! salarial_0,-2+o! salda_-5+pagamento,+dor!
salg_?0?+amento,?-1?+ador! salient_-1+cia,+ador! salin_0,-2! salomonic_0,+o! salso_-2,0! salte_+ada,+ador!
salubr_+idade,+e! saluta_0,-4+ude! salv_+amento,+ador! salvadorenh_0,-3! samba_-0,-1+ista!
sanar_!-5+cura,-1+dor! sande_-1+ice,+eu! sangr_+amento,+ador! sangra_+mento,+do! sanguine_0,-3+e!
sanitari_0,-6+ude! santo_-1+idade,0! sao_!-1+nidade,0! sapatea_+do,+dor! sapec_+a,+ador!
saquea_-6+roubo,+dor! saracote_+io,+ador! sarar_-5+cura,-1+do! sarcas_+mo,0! sarja_+dura,+dor!
satanic_0,+o! satisfa_+cao,+torio! sauda_+cao,+dor! saudav_0,-2+e! saudos_-2+ade,0! sebent_+ice,+o!
seca_+gem,+dor! seco_!-1+ura,0! secretari_+ado,+o! secreto_-2+do,0! sectari_0,-5+ita! secular_0,-2+o!
sedentari_+edade,+o! sedento_-3,0! sediar_!-1+cao,-3+e! sedicios_-3+ao,0! sedoso_0,-3+a! sedut_-1+cao,0!
segar_!-1,-1+dor! segar_-1,-1+dor! segreda_-1+o,+dor! segur_+anca,+ador! selar_-1+gem,-1+dor!
selecionav_-5+ao,0! selvage_+ria,+m! semanal_0,-1! semantic_0,+a! semea_+dura,+dor! semelhan_+ca,+te!
semi_0,0! semidivin_+dade,+o! seminal_0,-4+ente! seminaristic_0,-4+o! seminu_!+dez,0! semitic_-1+smo,0!
senhor_+ia,0! senhorial_0,-3! sensabor_+ia,0! sensacionalis_+mo,+ta! sensat_+ez,+o! sensorial_0,-5+acao!
sensual_+idade,0! senta_+da,+dor! senti_-2+sacao,-2+sivel! septic_+idade,+o! sepulcral_0,-2+o!
sepult_+amento,+ador! sequestra_-1+o,+dor! ser_!-3+ocorrencia,-3+ocorrente! sereno_-1+idade,0! serial_0,-2+e!
serio_-1+edade,0! serramar_0,0! serran_0,-1! serrot_+agem,+ador! sertan_0,-1+o! servent_+ia,+e!
servical_+ismo,0! servil_+ismo,0! servir_-1+co,-1+dor! servo_-1+idao,0! sestros_-1,0! setencia_-2+a,+dor!
setentrional_0,0! setenviral_0,-2+os! setorial_0,-3! sicari_0,-6+crime! sido_!-4+ocorrencia,0!
siga_-1+uimento,-1+uidor! sigilos_-1,0! siglistic_+a,+o! sigo_-1+uimento,-1+uidor! silen_+cio,+ciador!
silenci_+o,+oso! silvestre_!0,-8+elva! simetrico_-2+a,0! simpat_+ia,+izante! simples_-2+icidade,-1+s!
sinagogal_0,-1! sincicial_0,-2+o! sincipital_0,-4+ucio! sincopa_-1+e,+dor! sindicatori_0,-2!
sinematic_0,-9+estames! singel_+eza,+o! singr_+adura,+ador! sinodal_0,-2+o! sintatic_0,-3+xe! sintetic_0,-3+se!
sirg_+a,+ador! siringeal_0,-2! sismic_0,-6+terremoto! sistemic_0,+o! sisud_+ez,+o! sitiar_-2+o,-1+dor!
so_!+lidao,0! soar_-2+m,-1+nte! sobej_+idao,+ador! soberan_+ia,0! soberb_+a,0! sobra_-0,+do!
sobran_+caria,+ceiro! sobre_0,0! sobressai_-9+realce,+do! sobressalt_+o,+ador!
sobrevir_!-8+sucessao,-8+sucessor! sobrevoa_+da,+dor! sobrio_-1+edade,0! socar_-2+o,-1+dor!
social_0,-2+edade! socialis_+mo,+ta! societari_+ado,+io! socio_!-1+edade,0! socobr_+o,+ador!
socorr_+o,-6+ajudante! socratic_0,-2+es! soer_!-4+frequencia,0! sofism_+a,+ador! sofreg_+uidao,+o!
sofriv_-6+mediocridade,0! solap_+amento,+ador! solar_?0?-1+da,?-2?-1+dor! solda_+gem,+dor!
solene_-1+idade,0! soler_+cia,+te! solicito_-1+ude,0! solidari_+edade,+o! solinh_+o,+ador! solist_-3+o,+a!
solitario_-5+dao,0! solta_-1+ura,+dor! solv_+encia,+ente! somar_-1,-1+dor! somatic_0,-7+corpo!
sombr_+a,+eador! somitic_+aria,+o! sonambul_+ismo,+o! sonar_-1+ncia,-1+nte! sonda_+gem,+dor!
sonha_-1+o,+dor! sonif_0,-2+o! sonoro_-1+idade,0! sons_+ice,+o! sopes_+o,+ador! sopont_+adura,+ador!
sopor_-3+nolencia,0! sopr_+o,+ador! soprar_-6+assopramento,-1+dor! sordid_+ez,0! soropositiv_0,+o!
sororal_0,-2! sorri_+so,+dente! sortea_-1+io,+dor! sortu_-1+e,+do! sorv_+o,+edor! sosseg_+o,0!
soterop_0,-6+alvador! soub_-3+abedoria,-3+abio! sova_-0,+dor! sovel_+ada,+ador! sovietic_0,0! sovin_+ice,+a!
suar_!-1+douro,-1+dor! suave_-1+idade,0! subcontrari_+edade,+o! subdesenvolv_+imento,0! subir_-1+da,-1+do!
subjaze_-2+cencia,-1+ido! sublime_-1+idade,0! submerg_-1+sao,+ente! submete_-3+issao,+dor!
subordina_+cao,+do! subservi_+encia,+ente! subsidia_-1+o,+dor! subsol_+agem,+ador!

substantiv_?0?+acao,+ador! substit_+uicao,+uto! subversiv_-2+ao,0! sucede_-2+ssao,-2+ssor! sucedid_0,0!
sucinto_-7+concisao,0! sudeste_!0,0! sudoeste_!0,0! sueco_0,-1+ia! sugar_-1+cao,-1+dor! sugeri_-2+stao,+dor!
suicid_+io,+a! suico_0,-1+a! suino_0,-5+porco! suja_-1+eira,+dor! sujeitar_-3+cao,-1+dor! sujo_-1+eira,0!
sul_!0,0! suleste_!0,0! sulin_0,-2! sumaria_-1+o,+nte! sumir_-1+co,-1+dor! sumo_-1+idade,0! super_0,0!
superabund_+ancia,+ante! superflu_+idade,+o! superpode_-9+poder,0! supersti_+cao,+cioso! supervi_+sao,+sor!
suplementa_?+ridade?-1+o,+r! suplic_+a,+ador! suporif_0,-2! suporta_-7+resistencia,+dor! suprem_+acia,+o!
sural_0,-1! surd_+ez,+o! surpre_+sa,+endente! surr_+a,+ador! surreal_+ismo,0! surrealis_+mo,+ta!
surrib_+a,+ador! suseran_+ia,0! suspeit_+a,+ador! suspend_-1+sao,-1+sivo! suspens_+ao,+o!
suspira_-1+o,+dor! sussurr_+o,+ador! suste_+ntacao,+ntador! susto_-1+acao,-1+ador! sut_+amento,+ador!
sutil_+eza,0! tabelar_?0?-1,?-1?-1+dor! tabular_0,-3+a! tacanh_+ice,+o! tacar_-1+da,-1+dor! tacha_+cao,+dor!
tafulh_+o,+ador! talar_-1+mento,-1+dor! talh_+a,+ador! taling_+adura,+ador! talon_+agem,+eiro!
talud_+amento,+ador! tamanh_+o,0! tampar_-1+mento,-1+dor! tara_-0,+dor! tarda_+nca,+dor! tarifari_0,-2!
tartamude_+io,+ador! tartare_+io,+ador! tasc_+a,+ador! tatua_+gem,+dor! taxia_+mento,+dor! taxist_0,-2!
teca_-2+xtura,-2+celao! tece_-2+xtura,+lao! teci_-2+xtura,-1+elao! teclar_-1+da,-1+dista! tedios_-1,0!
teim_+osia,+oso! tele_0,0! telefona_-1+ema,+dor! telepat_+ia,+a! televisiv_-2+ao,0! telha_+dura,+dor!
teme_-1+or,+roso! temerar_-2+idade,+io! temper_+o,+ador! tempestuos_-3+ade,0! temporao_-1+neidade,0!
temporari_+edade,+o! tended_+ura,+or! tenistic_0,-3! tenr_+ura,+o! tente_+io,+ador! tenu_+idade,+e! ter_!0,0!
terebintin_+agem,+ador! termic_0,-6+calor! terminal_0,0! terminar_-2+o,-1+dor! terno_-1+ura,0! terraqueo_0,-4!
terren_0,-2+a! terrestre_0,-5+a! terrific_+ancia,+ante! territ_+orio,+orial! terro_+r,0! terros_0,-2+a!
teso_!-1+ura,0! testa_-1+e,+dor! testemunha_-0,0! testicond_+ia,0! testo_!-5+seriedade,0! tesud_-2+ao,0!
tetrarritm_+ia,+ico! tetric_+idade,+o! tetro_-1+icidade,0! teutonic_0,-8+alemanha! texan_0,-1+s! textural_0,-1!
tibial_0,-1! tibio_-1+eza,0! tic_-1+que,+ador! tigrad_0,-2+e! timbr_+agem,+ador! timpanal_0,-2+o! tini_+do,+dor!
tinj_-1+cao,-1+gidor! tinta_+agem,-5+entintador! tir_+ada,+ador! tiranic_-1+a,+o! tirar_-1+da,-1+dor!
tisn_+a,+ador! tissular_0,-7+ecido! titanic_0,+o! titubea_+da,+dor! toar_!-1+da,-1+dor! tocante_0,0!
todo_!-2+talidade,0! told_+a,+ador! tolera_+ncia,+nte! tolerant_+ismo,+e! tolo_-1+ice,0! tomar_-1+da,-1+dor!
tonitru_+ancia,+ante! tonsilar_0,-1! tonsur_+a,+ador! tonte_+io,+ador! tonto_-1+ura,0! topa_+da,+dor!
topico_!0,0! toqu_+e,-2+cador! torce_-1+ida,+dor! tormentori_0,-2! torna_+da,+dor! torpe_!+za,0!
torrencia_0,-3+te! torto_-1+uosidade,0! tortur_+a,+ador! tosa_-0,+dor! tosqui_+a,+ador! tossi_-1+e,+dor!
tossica_-2+dela,+dor! tosta_+dura,+dor! tourea_+cao,-1+iro! trabalh_+o,+ador! trabecular_0,-1!
tracar_-2+o,-1+dor! tradut_-1+cao,0! trafega_-1+o,+nte! trafic_+o,+ante! tragic_-2+edia,+o! trair_-1+cao,-1+dor!
traja_-1+e,+dor! tramar_-1+,-1+dor! tranquilo_-1+idade,0! trans_0,0! transar_-1,-1+dor! transatlantic_0,0!
transeunt_-9+caminho,+e! transfu_+sao,+sor! transgr_+essao,+essor! transig_+encia,+ente! transigiv_-2+encia,0!
transit_+o,+ador! translad_+o,+ador! transmis_+sao,+sor! transpar_+encia,+ente! transplanta_-1+e,+dor!
transporta_-1+e,+dor! traqueal_0,-1! traquin_+ice,+as! trasfeg_+a,+ador! traslad_+o,+ador!
tratar_-1+mento,-1+dor! traumatiza_-2+smo,+dor! travess_+ia,-7+atravessador! travesso_-1+ura,0!
traze_-2+nsporte,-2+nsportador! treina_-1+o,+do! tremer_-2+or,-1+dor! tremul_-2+or,+ante! trepa_+da,+dor!
trescal_+o,+ador! tresdobr_+adura,+ador! tresmalh_+o,+ador! trespass_+e,+ador! tresvari_+o,+ador!
tribal_0,-2+o! tributav_-1+cao,0! trigin_+ia,0! trigo_!0,0! trilh_+a,+ador! trinca_+dura,+dor! triplice_-1+idade,0!
tripudi_+o,+ador! trisc_+a,+ador! trist_+eza,+e! triunf_?0?+o,?+o?+ador! troa_+da,+dor! troca_-0,+dor!
troclear_0,-1! tromba_+da,+dor! trompa_+da,+dor! tronar_-1+da,0! tronch_+ura,+o! tropea_+da,-1+iro!
tropeca_-1+o,+nte! tuberculos_+e,0! tumoral_0,-2! tungar_-2+o,-1+dor! turbilhonar_0,-4+ao! turco_0,-2+quia!
tutelar_-1,-4+or! tutuc_-1+que,+ador! ubere_-1+dade,0! ubiqu_+idade,+o! ufan_+ia,0! uiva_-1+o,+dor!
ultraja_-1+e,+nte! ultrapass_+agem,+ante! umbilical_0,-5+go! umbratic_0,-8+sombra! umeral_0,-2+o!
umid_+ade,+o! ungi_-2+cao,+dor! ungir_-3+cao,-1+dor! ungueal_0,-5+ha! unicelular_0,0!
uniform_?+idade?+izacao,+izador! unilinear_0,-1! unimolecular_0,-1! unisson_+ancia,+o! universal_0,-2+o!
unta_-1+ura,+dor! urbano_-1+idade,0! ureteral_0,-2! uretral_0,-1! urinari_0,-2! uropigial_0,-2+o!
urticace_0,-4+ga! us_+o,+ador! usar_!-2+o,-2+uario! usufru_+to,+idor! uvular_0,-1! vacina_+cao,+dor!
vagabun_+dagem,+do! vagar_!-1,-1+nte! vaginal_0,-1! vago_!-0,0! vaiar_-1,-1+dor! valar_0,-1! valecular_0,-1!
valent_+ia,+e! valer_-5+avaliacao,-5+avaliado! valios_-2+a,0! valvar_0,-1! vampir_+ismo,+o! vandal_+ismo,+o!
vanta_+gem,+joso! vaqueja_+da,+dor! varao_-2+onia,0! variega_+cao,+dor! vario_-1+edade,0! varr_+ida,+edor!
vascain_0,-3+o! vascular_0,-5+o! vasculh_+ada,+ador! vasto_-1+idao,0! vaticin_+io,+ador!
vaziar_-6+esvaziamento,-2+ador! vedar_-1+cao,-1+dor! vegeta_+cao,+l! vegetal_0,0! vegetarian_+ismo,+o!
vegobscen_+idade,+o! veleja_-6+navegacao,+dor! velhac_+aria,+o! velho_-1+ice,0! venda_-5+cegamento,+dor!
vende_-1+a,+dor! venenos_-1,0! venta_-1+o,-1+o! ventral_0,-2+e! ventrilo_+quia,+co! venust_+idade,+o!
verbal_!+ismo,0! verdade_0,0! verde_0,0! verdin_0,-2+e! verga_+dura,+dor! vergonhos_-2+a,0!
veridic_0,-4+dade! vermelh_+idao,+o! vermelha_-1+o,+dor! vermelho_-1+idao,0! vermicular_0,-6+e!
vernacul_+idade,+o! verossimil_+hanca,0! versej_+adura,+ador! verter_-1+dura,-1+dor! vertig_0,+em!
vesical_0,-7+bexiga! vesicular_0,-1! vess_+adela,+ador! vesti_+menta,+dor! vestibulan_-1+r,0! vestibular_0,0!
vetar_-2+o,-1+nte! veteran_+ice,+o! vetust_+ez,+o! viajar_-3+gem,-1+nte! vibra_+cao,+dor! viciar_-2+o,-1+dor!
vicos_-1,+o! vidam_+ia,+a! vigen_+cia,+te! vigiar_-1,-1! vigil_+ancia,+ante! vigo_+r,+roso!
vigora_-3+encia,-3+ente! vil_!+eza,0! vila_+nia,+o! vinagr_+eza,+e! vindic_+ia,0! vinga_+nca,+dor! vini_0,-1+ho!
vinic_0,-2+ho! vinil_0,0! violar_-1+cao,-1+dor! violen_+cia,+tador! vir_!-1+nda,-1+ndo! viral_0,-2+us!
virar_-1+da,-1+dor! virgem_-2+indade,-1+m! virid_+encia,+ente! virtuo_-1+de,+so! visar_-1+da,-1+dor!
visceral_0,-1+s! vistos_0,+o! visual_0,0! vital_0,0! vitalici_+edade,+o! vitamina_-0,origina_-3+em,+dor!
vitic_0,-3+nho! vitima_-0,+dor! vitre_0,-3+dro! viuvo_-1+ez,0! vive_-2+da,-1+o! vivencia_-0,+dor! vivo_-2+da,0!
viza_-2+sto,+dor1 vizinh_+anca,+o! voa_-1+o,+dor! vocal_0,-3+z! vogar_!-5+navegacao,-1+dor! volt_+a,+ador!
volumo_-1+e,+so! voluntari_+edade,+o! volve_-4+olta,+dor! vomit_+o,+ador! vota_-1+o,+nte! votivo_0,-3+o!
voze_+aria,+ador! vulvar_0,-1! warrant_+agem,+ador! xucr_+ice,+o! zabumb_+ada,+ador! zanga_-0,+dor!
zanguizarre_+io,+ador! zebral_0,-1! zebuin_0,-2! zelar_-2+o,-1+dor! zelos_-1,0! zigezaguea_-1,+nte!
zizi_+amento,+ador! zoa_+da,+dor! zoar_!-1+da,-1+dor! zomba_+ria,+dor! zonal_0,-1! zoofil_+ia,0!
zumbi_+do,+dor! zuni_+do,+dor! zurz_+idela,+idor!

Autômato para padrões de adjetivos

abitado_-2+cao,0! abrido_-2+ertura,0! acado_-2+mento,0! acante_-3,0! acendido_-3+imento,0! acido_-1+ez,0!
acoso_-2,0! actado_-2+cao,0! adido_-4+sao,0! ado_-2+cao,0! ador_-3+cao,-0! afado_-2+mento,0!
agado_-2+mento,0! agador_-3+mento,-0! ageado_-3+m,0! agido_-4+cao,0! aico_0,-4+eu! ajoso_-4+gem,0!
al_+idade,0! alado_-2+cao,0! algado_-2+da,0! alido_-1+ez,0! alizado_-2+cao,0! almado_-2+mento,0!
amado_-2+mento,0! amico_0,-1+a! amorado_-3+o,0! anado_-2+mento,0! ancoso_-3+a,0! andido_-5+sao,0!
anico_-2+smo,0! anoso_-2,0! apado_-2+mento,0! aprendido_-3+izado,0! apturado_-3+a,0! ar_+idade,0!
arado_-2+cao,0! ario_-1+smo,0! artado_-3+e,0! asado_-2+mento,0! asador_-3+mento,-0! asiado_-3+a,0!
asmico_0,-3+a! assado_-2+gem,0! assador_-3+gem,-0! astico_-1+a,0! atado_-2+mento,0! atador_-3+mento,-0!
atante_-3+cao,0! atario_0,0! atento_-2+cao,0! atico_-2+smo,0! ativo_-4+cao,0! ator_-3+cao,-0!
aurado_-2+cao,0! avado_-2+mento,0! avisado_-3+o,0! avisador_-4+o,-0! axado_-2+mento,0! az_-1+cia,0!
azado_-2+mento,0! balizado_-2+mento,0! batido_-3+e,0! berante_-2+cia,0! bico_0,0! biliario_0,-8+vel!
bilitado_-4+dade,0! borado_-2+cao,0! brado_-2+mento,0! brante_-3+cao,0! buido_-2+cao,0! cabado_-2+mento,0!
cadista_0,-4+o! caido_-2+da,0! calibrado_-2+gem,0! canceroso_-3,0! cante_-3+cao,0! cavado_-2+cao,0!
cefalo_-1+ia,0! centrico_-2+smo,0! certado_-3+o,0! certante_-4+o,0! certo_-1+eza,0! cessario_-4+idade,0!
chado_-2+mento,0! ciado_-2+mento,0! ciador_-3+mento,-0! ciario_0,-3! cidado_-2+cao,0! cidido_-3+encia,0!
cifrado_-2+cao,0! cinado_-3+io,0! cionado_-6+ao,0! cional_0,-5+ao! cipado_-2+cao,0! citado_-2+cao,0!
cluso_-1+ao,0! colonial_0,-1! comico_-1+idade,0! comodo_-1+idade,0! conexo_-1+ao,0! conjurado_-2+cao,0!
corado_-2+cao,0! cotico_-4+se,0! crado_-2+cao,0! crata_-2+cia,0! cratico_-4+cia,0! crente_-2+ca,0!
criado_-2+cao,0! cromatico_-2+smo,0! ctivo_-4+cao,0! ctual_0,0! culto_-1+ura,0! culturado_-3+a,0!
cundo_-1+idade,0! cupado_-2+cao,0! curado_-3+a,0! custico_0,-1+a! dado_-2+mento,0! dante_-3+mento,0!
dario_0,0! daz_-1+cidade,0! decente_-2+cia,0! denado_-2+cao,0! dental_0,-2+e! desco_0,0! desimo_0,0!
diario_0,-4+o! dicado_-2+cao,0! dico_0,0! didatico_0,-1+a! dido_-2+cao,0! dioptrico_-1+a,0! disciplinar_0,-1!
distado_-2+ncia,0! distante_-2+cia,0! ditado_-4+cao,0! dizido_-4+cao,0! doente_-2+ca,0! dorado_-2+cao,0!
drado_-2+mento,0! dresco_0,-4! durado_-2+cao,0! dutor_-3+cao,-0! duzido_-4+cao,0! eado_-2+mento,0!
eante_-3+mento,0! ebado_-2+mento,0! ecado_-2+mento,0! ecente_-4+imento,0! ecidido_-4+sao,0!
ectado_-4+cao,0! ecutivo_0,0! edido_-2+encia,0! edor_-4+imento,-0! edulo_-1+idade,0! eendido_-4+so,0!
egado_-2+mento,0! eguro_-1+anca,0! eico_0,-2+a! eiro_0,0! ejado_-3+o,0! eletrico_-1+idade,0!
eletronico_0,-1+a! elido_-2+mento,0! embrado_-2+nca,0! emico_0,-2+a! emorado_-2+cao,0! empregado_-3+o,0!
enado_-2+mento,0! enal_0,0! ene_-1+idade,0! enido_-3+cao,0! entado_-2+cao,0! entendido_-3+imento,0!
ento_-2+cia,0! entrado_-2+da,0! equilibrado_-3+io,0! erado_-2+cao,0! erante_-3+cao,0! erico_-2+smo,0!
erido_-3+encia,0! erioso_-2,0! ero_-1+idade,0! ersado_-2+o,0! ersor_-2+ao,-0! ertido_-4+sao,0!
esado_-2+mento,0! esador_-3+mento,-0! esco_0,-4+o! esente_-2+ca,0! esico_-2+a,0! estido_-2+mento,0!
estrutivo_-4+icao,0! etado_-2+mento,0! etante_-3+cao,0! etico_-2+smo,0! etitivo_-1+idade,0! etnico_0,-2+a!
etrado_-2+cao,0! eutico_-6+ia,0! evado_-2+mento,0! exado_-2+cao,0! exo_-1+idade,0! ezado_-2+mento,0!
fadado_-3+o,0! fago_-1+ia,0! ferico_0,-3+a! ferido_-4+ncia,0! ferivel_-4+encia,0! fertado_-3+a,0!
fessado_-6+issao,0! fessor_-5+issao,-0! festado_-2+cao,0! fetado_-2+cao,0! fiado_-2+mento,0! ficado_-2+cao,0!
ficavel_-3+cao,0! fico_-2+a,0! fiel_-2+delidade,0! filtrado_-2+cao,0! finito_-1+ude,0! fisico_0,-1+a!
fixado_-2+cao,0! flado_-2+cao,0! fluido_-3+encia,0! fobico_-2+a,0! fobo_-1+ia,0! fonico_-2+a,0!
friado_-2+mento,0! furado_-2+cao,0! fuso_-1+ao,0! gante_-3+cao,0! gatado_-3+e,0! gaz_-1+cidade,0!
genico_-2+a,0! gerido_-4+stao,0! gestado_-2+cao,0! gestivo_-3+ao,0! gico_-2+a,0! gitado_-2+cao,0!
gnado_-2+cao,0! gonal_0,0! gorante_-3+mento,0! gostado_-3+o,0! governado_-3+o,0! grado_-2+cao,0!
grafado_-3+ia,0! grante_-3+cao,0! grato_-1+idao,0! gregado_-2+cao,0! guado_-2+mento,0! guaio_0,-1!
guinte_-3+mento,0! gulhado_-3+o,0! gurado_-2+cao,0! gustado_-2+cao,0! hacado_-3+o,0! hante_-3+cao,0!
iante_0,0! ibido_-2+cao,0! icado_-2+mento,0! icio_0,0! ico_-1+idade,0! idido_-4+sao,0! ido_-2+mento,0!
idoso_-3+ade,0! iedoso_-3+ade,0! iental_0,-2+e! igado_-2+mento,0! igno_-1+idade,0! igual_+dade,0!
il_+idade,0! ilado_-2+cao,0! ilizado_-4+dade,0! ilmado_-2+gem,0! ime_-1+idade,0! imente_-6+essao,0!
imetrico_0,-3+o! imo_-1+idade,0! inado_-2+mento,0! inante_-3+cao,0! indido_-5+sao,0! industrial_0,-1!
ingido_-2+mento,0! inido_-2+cao,0! ioante_-6+ao,0! ionante_-7+ao,0! ionario_-7+ao,0! iorado_-2+cao,0!
ipado_-2+mento,0! iquido_-1+ez,0! irado_-2+mento,0! irico_-2+smo,0! irmado_-2+cao,0! iso_-1+ao,0!
issimo_0,0! ista_0,0! istico_-4+mo,0! itado_-2+mento,0! italo_0,-1+ia! itante_-3+cao,0! itativo_-6+dade,0!
itido_-4+ssao,0! itivo_-4+cao,0! itual_0,-3+o! ivado_-2+mento,0! ivante_-3+cao,0! ivo_-1+idade,0!
ivoco_-1+idade,0! ixado_-2+mento,0! izado_-2+mento,0! izante_-3+cao,0! jado_-2+mento,0! jador_-3+mento,-0!
jante_-3+mento,0! jetado_-4+cao,0! jugado_-2+cao,0! justado_-2+mento,0! justo_-1+ica,0! lado_-2+mento,0!
lantado_-2+cao,0! lante_0,0! lastico_-1+idade,0! lcado_-2+mento,0! leal_+dade,0! legado_-2+cao,0!
lentador_-4+o,-0! lesco_0,-4! leto_-2+cao,0! levado_-2+cao,0! lgado_-2+mento,0! lgido_-3+encia,0!
lhado_-2+mento,0! liberal_+ismo,0! licado_-2+cao,0! lico_0,0! lidado_-2+cao,0! linado_-2+cao,0! lionario_0,0!
litado_-2+cao,0! litico_-4+se,0! locado_-2+cao,0! logo_-1+ia,0! logrado_-3+o,0! lorado_-2+cao,0!
lotado_-2+cao,0! lpado_-2+mento,0! lsivo_-3+ao,0! ltuado_-4+o,0! luido_-3+sao,0! maniaco_-2,0!
marcado_-2+cao,0! mario_0,0! matico_0,-4! mbarcado_-4+que,0! mbido_-4+ncia,0! mbinado_-2+cao,0!
mbrante_-3+mento,0! medido_-2+cao,0! mental_0,-2+o! mentario_0,-4+o! mergido_-4+sao,0! merico_0,-3+o!
mero_0,0! mestral_0,-2+e! metico_-1+a,0! metido_-3+imento,0! metrico_-2+a,0! minado_-2+cao,0!
missado_-3+o,+do! misso_-1+ao,0! mitado_-2+cao,0! mocado_-3+o,0! moido_-3+ida,0! molado_-2+cao,0!
monial_0,-2+o! monico_-2+a,0! montado_-2+gem,0! movido_-4+cao,0! munido_-2+cao,0! nario_0,0!
ncado_-2+mento,0! ncial_0,-1! ndido_-4+sao,0! nectado_-5+xao,0! negado_-2+cao,0! neo_-1+idade,0!
nesco_0,-5+o! nestado_-2+cao,0! netario_0,-3! ngido_-3+encia,0! nhado_-2+mento,0! nhoso_-2,0!
nioso_-3+a,0! nistrado_-2+cao,0! nizado_-2+cao,0! nolvido_-3+imento,0! notado_-2+cao,0! nsado_-2+mento,0!
nsador_-3+mento,-0! nsioso_-4+ao,0! nso_-1+ao,0! ntado_-2+mento,0! ntario_0,0! nte_-2+cia,0! ntivo_-4+cao,0!
nuante_-3+cao,0! nunciado_-3+o,0! oado_-2+mento,0! obedecido_-5+iencia,0! oberto_-4+rimento,0!
ocado_-2+mento,0! ociado_-2+cao,0! odido_-4+sao,0! odoxo_-1+ia,0! odutivo_-1+idade,0! ofilo_-1+ia,0!
olido_-1+ez,0! olorido_-6+r,0! olsado_-3+o,0! olsador_-4+o,-0! oltado_-3+a,0! olvido_-4+ucao,0!
ombado_-2+mento,0! omecado_-3+o,0! omico_0,-2+a! ominavel_-3+cao,0! ompido_-3+imento,0!
onado_-2+mento,0! onesto_-1+idade,0! ongado_-2+mento,0! onho_-4+eza,0! onimo_-1+ia,0! ono_-1+ia,0!

onrado_-3+a,0! ontador_-3+mento,-0! ontal_0,-2+e! ontrado_-3+o,0! onunciado_-3+a,0! ope_-1+ia,0!
opico_-2+a,0! oplastico_-2+a,0! oprio_-1+edade,0! orado_-2+mento,0! orescido_-2+encia,0! orfico_-2+smo,0!
oriado_-3+a,0! orial_0,-2+o! orico_0,-2+a! orio_-1+edade,0! orioso_-3+a,0! ormado_-2+cao,0! ornado_-3+o,0!
oroso_-3,0! orrado_-2+cao,0! orrido_-6+urso,0! ortado_-3+e,0! ortador_-4+e,-0! orvido_-4+cao,0! osivo_-3+ao,0!
oso_-1+idade,0! osturado_-3+a,0! otado_-2+mento,0! otante_-3+cao,0! otico_-2+smo,0! otivo_-4+cao,0!
ovante_-3+cao,0! ovente_-4+imento,0! pactuado_-4+o,0! pante_-3+cao,0! pargido_-2+mento,0!
partido_-2+cao,0! patico_-2+a,0! paz_-1+cidade,0! pedido_-2+mento,0! pergido_-4+sao,0! pertado_-3+o,0!
pesco_0,-3! piado_-3+o,0! picado_-2+da,0! pidado_-2+cao,0! pidante_-3+cao,0! pido_-1+ez,0! pirado_-2+cao,0!
pitado_-2+cao,0! pitante_-2+cia,0! pleto_-1+ude,0! plinado_-2,0! plo_-1+icidade,0! polado_-2+cao,0!
politico_-1+a,0! ponente_-5+sicao,0! porado_-2+cao,0! portado_-2+cao,0! positado_-3+o,0! prado_-2+mento,0!
prazido_-2+r,0! preciavel_-3+cao,0! pressivo_-3+ao,0! prestado_-3+imo,0! pretado_-2+cao,0!
primido_-5+essao,0! protegido_-4+cao,0! provido_-3+imento,0! ptado_-2+cao,0! pudico_-1+ia,0!
purado_-2+cao,0! quico_0,-2+a! quieto_-1+ude,0! quimico_-1+a,0! quistado_-3+a,0! radico_-1+idade,0!
rario_0,0! ratado_-2+cao,0! raz_-1+cidade,0! rcado_-2+mento,0! rciado_-3+o,0! redido_-4+ssao,0!
regrado_-2+mento,0! renado_-2+gem,0! residencial_0,-1! revido_-2+sao,0! rfeito_-2+cao,0! rfista_0,-4+e!
rgado_-2+mento,0! rgido_-3+encia,0! riado_-2+mento,0! ricado_-2+cao,0! rigado_-2+cao,0! rior_+idade,-0!
riscado_-2+mento,0! ritado_-2+gem,0! rivado_-2+cao,0! rizado_-2+cao,0! rmado_-2+mento,0! rme_-1+idade,0!
rno_-1+idade,0! rotado_-2+cao,0! rrado_-2+mento,0! rreto_-2+cao,0! rulhado_-3+o,0! rupto_-2+cao,0!
rustado_-2+cao,0! rustrado_-2+cao,0! saltado_-3+o,0! sante_0,0! scido_-3+imento,0! screvido_-5+icao,0!
scrito_-2+cao,0! sedado_-2+cao,0! sidido_-3+encia,0! sindical_0,-1+to! sionado_-6+ao,0! sional_0,-5+ao!
sitado_-2+cao,0! sorio_-4+ao,0! speitado_-3+o,0! speito_-2+cao,0! ssado_-2+mento,0! ssador_-3+mento,-0!
sseado_-3+io,0! ssico_0,0! ssional_0,-8+issao! ssivo_-3+ao,0! stado_-2+mento,0! stante_0,0! stico_-4+mo,0!
stido_-3+encia,0! stinto_-2+cao,0! sual_0,-3+o! sumido_-4+ncao,0! surado_-3+a,0! sustado_-2+cao,0!
tacado_-4+que,0! talhado_-3+e,0! tario_0,-5+dade! tatico_-1+a,0! taxado_-2+cao,0! tecnico_-1+a,0!
tenso_-1+ao,0! terapico_-2+a,0! ternal_0,-2+idade! tesco_0,-3! tetico_-1+a,0! tico_-2+a,0! tido_-3+encao,0!
tigado_-2+cao,0! tilado_-2+cao,0! tituido_-2+cao,0! tivado_-2+cao,0! tizado_-2+cao,0! tocado_-4+que,0!
torcido_-3+ao,0! torio_-5+cao,0! toso_-2,0! trado_-2+mento,0! traente_-4+cao,0! traido_-3+cao,0! tral_0,-2+o!
trante_-3+cao,0! trato_-2+cao,0! trico_-2+a,0! trito_-2+cao,0! tropical_+ismo,0! tuante_-3+cao,0!
tuinte_-3+cao,0! turado_-2+cao,0! tutivo_-4+icao,0! ucado_-2+mento,0! ucido_-1+ez,0! udido_-4+sao,0!
ugado_-2+mento,0! uistico_-1+a,0! ulado_-2+cao,0! ultado_-2+cao,0! ultimo_0,0! umano_-1+idade,0!
undido_-5+sao,0! une_-1+idade,0! unhado_-3+o,0! unido_-2+ao,0! uo_-1+idade,0! urado_-2+mento,0!
urdido_-2+mento,0! ursado_-3+o,0! ursador_-4+o,-0! urtado_-2+mento,0! uscado_-2+mento,0! usivo_-3+ao,0!
ustico_-1+idade,0! usuario_0,0! utante_-3+cao,0! utario_0,-4+o! utico_-1+a,0! utivo_-4+cao,0! uvante_-3+cao,0!
uzado_-2+mento,0! uzido_-4+cao,0! va_-1+idade,0! valido_-1+ez,0! variavel_-3+cao,0! vavel_-5+babilidade,0!
vel_-3+bilidade,0! vendido_-3+a,0! vertido_-4+sao,0! viado_-3+o,0! viario_0,-3! vido_-2+sao,0! visado_-2+o,0!
visitado_-3+a,0! visor_-2+ao,-0! vistado_-3+a,0! vitado_-2+cao,0! vivido_-2+encia,0! vizado_-4+dade,0!
vocado_-2+cao,0! xpedido_-2+cao,0! xtual_0,-3+o! z_-1+cidade,0!

Autômato para padrões de verbos

(a cadeia s de cada entrada não contém, neste autômato, o “r” final)

a_+cao,+dor! abita_+cao,+dor! abri_-2+ertura,+dor! aca_+mento,+dor! acende_-1+imento,+nte! acta_+cao,+dor!
adi_-2+sao,-2+sor! afa_+mento,+dor! aga_+mento,+dor! agea_-1+m,-2+ista! agi_-2+cao,-1+ente! ala_+cao,+dor!
alga_+da,+dor! aliza_+cao,+dor! alma_+mento,+dor! ama_+cao,+dor! amora_-1+o,+dor! ana_+mento,+dor!
apa_+mento,+dor! aprende_-1+izado,-1+iz! aptura_-1+a,+dor! ara_+cao,+dor! arta_-1+e,+dor! asa_+mento,+dor!
asia_-1+a,+dor! assa_+gem,+dor! ata_+mento,+dor! aura_+cao,+dor! ava_+mento,+dor! avisa_-1+o,+dor!
axa_+mento,+dor! aza_+mento,+dor! baliza_+mento,+dor! bate_-1+e,+dor! bilita_-2+dade,+dor! bora_+cao,+dor!
bra_+mento,+dor! bui_+cao,+dor! caba_+mento,+dor! cai_+da,+dor! calibra_+gem,+dor! cava_+cao,+dor!
ceja_-1+o,+dor! centra_+da,+dor! certa_-1+o,+dor! cha_+mento,+dor! cia_+mento,+dor! cida_+cao,+dor!
cidi_-1+encia,-1+ente! cifra_+cao,+dor! cina_-1+io,+dor! ciona_-4+ao,+dor! cipa_+cao,+dor! cita_+cao,+dor!
conjura_+cao,+dor! cora_+cao,+dor! cra_+cao,+dor! cre_+nca,+nte! cria_+cao,+dor! cultura_-1+a,+dor!
cupa_+cao,+dor! cura_-1+a,+dor! da_+mento,+dor! deja_-1+o,+dor! dena_+cao,+dor! di_+cao,+dor!
dica_+cao,+dor! dista_+ncia,+nte! dita_-2+cao,-1+or! dize_-2+cao,+nte! dora_+cao,+dor! dra_+mento,+dor!
dura_+cao,+dor! duzi_-2+cao,-2+tor! e_-1+imento,+dor! ea_+mento,+dor! eba_+mento,+dor! eca_+mento,+dor!
ecidi_-2+sao,-2+sor! ecta_-2+cao,+dor! ede_+ncia,+nte! eende_-2+so,+nte! ega_+mento,+dor! eli_+mento,+dor!
embra_+nca,+dor! emora_+cao,+dor! emprega_-1+o,+dor! ena_+mento,+dor! eni_-1+cao,+dor! enta_+cao,+dor!
entende_-1+imento,+dor! entra_+da,+nte! envolve_-1+imento,+dor! equilibra_-1+io,+dor! era_+cao,+dor!
eri_-1+encia,+dor! ersa_+o,-1+or! erte_-2+sao,-2+sor! esa_+mento,+dor! esti_+mento,+dor! eta_+cao,+dor!
etra_+cao,+nte! eva_+mento,+dor! exa_+cao,+dor! eza_+mento,+dor! fada_-1+o,+dor! fatura_+mento,+dor!
faze_-2+cao,-2+tor! ferta_-1+a,+dor! fessa_-4+issao,-1+or! festa_+cao,+nte! feta_+cao,+dor! fia_+mento,+dor!
fica_+cao,+dor! filtra_+cao,+dor! fixa_+cao,+dor! fla_+cao,+dor! flui_-1+encia,-1+ente! fria_+mento,+dor!
fura_+cao,+dor! gata_-1+e,+dor! geri_-2+stao,-2+stor! gesta_+cao,+nte! gita_+cao,+dor! gna_+cao,+dor!
gosta_-1+o,+dor! governa_-1+o,+dor! gra_+cao,+dor! grafa_-1+ia,-1+o! grega_+cao,+dor! gua_+mento,+dor!
gulha_-1+o,+dor! gura_+cao,+dor! gusta_+cao,+dor! haca_-1+o,+dor! i_+mento,+dor! ibi_+cao,+dor!
ica_+mento,+dor! idi_-2+sao,-2+sor! iga_+mento,+dor! igi_-1+encia,-1+ente! ila_+cao,+dor! iliza_-2+dade,+dor!
ilma_+gem,+dor! ina_+mento,+dor! ingi_+mento,+dor! ini_+cao,+dor! iora_+cao,+dor! ipa_+mento,+dor!
ira_+mento,+dor! irma_+cao,+dor! ita_+mento,+dor! iti_-2+ssao,-2+ssor! iva_+mento,+dor! ixa_+mento,+dor!
iza_+mento,+dor! ja_+mento,+dor! jeta_-2+cao,+dor! juga_+cao,+dor! justa_+mento,+dor! la_+mento,+dor!
lanta_+cao,+dor! lca_+mento,+dor! lega_+cao,+dor! leva_+cao,+dor! lga_+mento,+dor! lgi_-1+encia,-1+ente!
lha_+mento,+dor! lica_+cao,+dor! lida_+cao,+dor! lina_+cao,+dor! lita_+cao,+dor! loca_+cao,+dor!
logra_-1+o,+dor! lora_+cao,+dor! lori_-4+r,+dor! lota_+cao,+dor! lpa_+mento,+dor! ltua_-2+o,+dor!
lui_-1+sao,+dor! marca_+cao,+dor! mbarca_-2+que,+dor! mbi_-1+encia,+dor! mbina_+cao,+dor!

medi_+cao,+dor! medita_+cao,+dor! mergi_-2+sao,-2+sor! mete_-1+imento,+dor! mina_+cao,+dor!
missa_-1+o,+do! mita_+cao,+dor! miza_+cao,+dor! moca_-1+o,+dor! moe_-1+ida,+dor! mola_+cao,+dor!
monta_+gem,+dor! move_-2+cao,+dor! muni_+cao,+dor! nca_+mento,+dor! nde_+ncia,+nte! ndi_-3+sao,+dor!
necta_-3+xao,+dor! nega_+cao,+dor! nesta_+cao,+dor! ngi_-1+encia,-1+ente! nha_+mento,+dor!
nistra_+cao,+dor! niza_+cao,+dor! nolve_-1+imento,+dor! nota_+cao,+dor! nsa_+mento,+dor! nta_+mento,+dor!
nuncia_-1+o,+nte! o_+sicao,+sitor! oa_+mento,+dor! obedece_-3+iencia,-3+iente! oca_+mento,+dor!
ocia_+cao,+dor! odi_-2+sao,+dor! oli_+cao,+dor! oliza_+cao,+dor! olsa_-1+o,+dor! olta_-1+a,+dor!
olve_-2+ucao,+dor! omba_+mento,+dor! omeca_-1+o,+dor! ompe_-1+imento,+dor! ona_+mento,+dor!
onga_+mento,+dor! onra_-1+a,+do! ontra_-1+o,+dor! onuncia_-1+a,+nte! ora_+mento,+dor! oresce_+ncia,+nte!
oria_-1+a,+dor! orma_+cao,+dor! orna_-1+o,+dor! orra_+cao,+dor! orre_-4+urso,+nte! orta_-1+e,+dor!
orve_-2+cao,+nte! ostura_-1+a,+dor! ota_+mento,+dor! pactua_-2+o,+nte! pargi_+mento,+dor! parti_+cao,+dor!
pedi_+mento,+dor! pergi_-2+sao,-2+sor! perta_-1+o,+dor! pia_-1+o,+dor! pica_+da,+dor! pida_+cao,+dor!
pira_+cao,+dor! pita_+cao,+dor! plina_-1+a,+dor! pola_+cao,+dor! pora_+cao,+dor! porta_+cao,+dor!
posita_-1+o,+nte! pra_+mento,+dor! praze_+r,+roso! presta_-1+imo,+dor! preta_+cao,+dor!
primi_-3+essao,-3+essor! protege_-2+cao,-2+tor! prove_-1+imento,+dor! pta_+cao,+dor! pura_+cao,+dor!
quista_-1+a,+dor! rata_+cao,+dor! rca_+mento,+dor! rcia_-1+o,+dor! redi_-2+ssao,-2+ssor! regra_+mento,+dor!
rena_+gem,+dor! reve_-1+isao,-1+isor! rga_+mento,+dor! rgi_-1+encia,-1+ente! ria_+mento,+dor!
rica_+cao,+dor! riga_+cao,+dor! rina_+cao,+dor! risca_+mento,+dor! rita_+gem,+dor! riva_+cao,+dor!
riza_+cao,+dor! rma_+mento,+dor! rota_+cao,+dor! rra_+mento,+dor! rulha_-1+o,+dor! rusta_+cao,+dor!
rustra_+cao,+dor! salta_-1+o,+dor! sce_-1+imento,+nte! screve_-3+icao,-3+itor! seda_+cao,+dor!
seja_-1+o,+dor! sidi_-1+encia,-1+ente! siona_-4+ao,+dor! sita_+cao,+dor! speita_-1+o,+dor! ssa_+mento,+dor!
ssea_-1+io,+dor! sta_+mento,+dor! sti_-1+encia,-1+ente! sumi_-2+ncao,+dor! sura_-1+a,+dor! susta_+cao,+dor!
taca_-2+que,+nte! talha_-1+e,+dor! taxa_+cao,+dor! te_+ncao,+ntor! teja_-1+o,+dor! tiga_+cao,+dor!
tila_+cao,+dor! titui_+cao,+dor! tiva_+cao,+dor! tiza_+cao,+dor! toca_-2+que,+dor! torce_-1+ao,+dor!
tra_+mento,+dor! trai_-1+cao,+dor! tura_+cao,+dor! uca_+mento,+dor! udi_-2+sao,-2+sor! uga_+mento,+dor!
ula_+cao,+dor! ulta_+cao,+dor! unha_-1+o,+dor! uni_+ao,+ficador! ura_+mento,+dor! urdi_+mento,+dor!
ursa_-1+o,+dor! urta_+mento,+dor! usca_+mento,+dor! uza_+mento,+dor! uzi_-2+cao,-2+tor! vale_+ncia,+nte!
ve_-1+isao,-1+isor! vende_-1+a,+dor! verte_-2+sao,-2+sor! via_-1+o,+dor! visa_+o,-1+or! visita_-1+a,+nte!
vista_-1+a,+dor! vita_+cao,+dor! vive_+ncia,+nte! viza_-2+dade,+dor! voca_+cao,+dor! xpedi_+cao,+dor!

Autômato para tratamento de sinonímia de substantivos

abafadela_!-4+mento,0! abaixadela_!-4+mento,0! abanadela_!-4+cao,0! abasto_!-1+amento,0! abate_!-
1+imento,0! abotoacao_!-3+mento,0! abridela_!-6+ertura,0! absorvedor_!0,-3+nte! acertada_!-3+o,0! acessao_!-
4+dencia,0! achatadela_!-4+mento,0! achegamento_!-6+o,0! acostagem_!-3+mento,0! acosto_!-1+amento!
acrescencia_!--3+tamento,0! acridade_!-3+ez,0! acuidade_!-7+gudeza,0! adstricao_!-3+ngencia,0! adubagem_!-
3+cao,0! afagamento_!-6+o,0! afeicoamento_!-7+ao,0! afligimento_!-7+cao,0! afogadela_!-4+mento,0! aforro_!-
1+amento,0! agarra_!+mento,0! agenciacao_!-3+mento,0! aguardamento_!-6+o,0! aguilhoadela_!-4+mento,0!
ajudador_!0,-3+nte! ajustamento_!-5+gem,0! alucinamento_!-5+cao,0! alumiacao_!-9+iluminacao,0! aluminaca_-
9+iluminacao,0! amansadela_!-4+mento,0! amarelidao_!-4+o,0! amassadura_!-4+mento,0! amassadela_!-
4+mento,0! amplidao_!-3+tude,0! andanca_!-3+mento,0! anunciacao_!-4+o,0! apalpadela_!-4+cao,0! aparo_!-
1+agem,0! aparadela_!-4+gem,0! apelacao_!-4+o,0! apertadela_!5+o,0! aprendizagem_!-3+do,0! aprendiz_!-
7+luno,0! aprontame_-9+prontidao,0! aracao_!-6+lavragem,0! aradura_!-7+lavragem,0! arador_!0,-6+lavrador!
ardencia_!-5+or,0! ardume_!-3+or,0! armacao_!-3+mento,0! armazenagem_!-3+mento,0! arrancadela_!-
4+mento,0! arranjamento_!-5+o,0! arrebentamento_!-5+cao,0! arrematacao_!-4+e,0! arrematante_!0,-3+dor,0!
arruiname_-9+ruina,0! assado_!-1+ura,0! assombracao_!-4+o,0! atracadela_!-4+cao,0! atropelo_!-1+amento,0!
barbeiragem_!-4+ismo,0! barradela_!-4+cao,0! batedura_!-5+ida,0! beberrao_!0,-5+ado! beberronia_!-6+deira,0!
bolina_!+cao,0! bolinagem_!-3+cao,0! bombeacao_!-3+mento,0! bonomia_!-4+dade,0! borradura_!-3+a,0!
brilhantura_!-6+o,0! brunidura_!-4+mento,0! caiadela_!-4+cao,0! caimento_!-8+queda,0! calcamento_!-7+que,0!
calcadura_!-6+que,0! calibracao_!-3+gem,0! cantadela_!-5+oria,0! cantador_!0,-4+or! capinacao_!-3,0!
captagem_!-3+cao,0! carda_!+cao,0! catacao_!-3,0! cavadela_!-4+cao,0! chamusca_!+mento,0! chupadela_!-
3+a,0! circunspecao_!-3+ccao,0! claridade_!-5+eza,0! coalhadura_!-3+a,0! coccao_!-4+zimento,0!
cozedura_!-5+imento,0! cochilo_!-1+ada,0! colacao_!-3+gem,0! complemento_!-1+acao,0! concebimento_!-
7+pcao,0! confortante_!-3+dor,0! confrontacao_!-4+o,0! confrontante_!0,-3+dor,0! consumicao_!-4+o,0! costeio_!-
2+agem,0! cravamento_!-5+cao,0! crestamento_!-5,0! cretinismo_!-3+ce,0! curticao_!-3+mento,0! dador_!-
4+oador,0! deposicao_!-5+imento,0! desdouramento_!-6+o,0! desencaixamento_!-6+e,0! desmedra_!+mento,0!
despistamento_!-6+e,0! desvelamento_!-6+o,0! devotamento_!-7+cao,0! dobradura_!-4+mento,0! dobre_!-
1+amento,0! domacao_!-3,0! dotamento_!-5+cao,0! douracao_!-3+mento,0! drama_!+ticidade,0! embebicao_!-
3+mento,0! embocadura_!-4+mento,0! empaste_!-1+amento,0! encabulacao_!-3+mento,0! encaixamento_!-6+e,0!
encanacao_!-3+mento,0! enchecao_!-4+imento,0! encostadela_!-4+mento,0! encruzamento_!-5,0! enformacao_!-
3+gem,0! engomadela_!-3+ura,0! engraxadela_!-4+mento,0! enlace_!-1+amento,0! enrascadela_!-2+a,0!
enredamento_!-6+o,0! enrolao_!0,-1+dor,0! ensaboadela_!-3+a,0! enterrada_!-2+mento,0! enterro_!-1+amento,0!
entristec_-9+tristeza,0! envasilhamento_!-9+amento,0! enxaguadura_!-5+e,0! enxertadura_!-5+o,0! equidade_!-
7+gualdade,0! escaldadura_!-4,0! escapa_!-1+e,0! escapulida_!-5+e,0! escoacao_!-3+mento,0! escorregadela_!-
4+mento,0! escovadela_!-3+a,0! escusacao_!-3,0! esfacelo_!-1+amento,0! esfregadela_!-3+a,0! esgotadura_!-
4+mento,0! esguichad_!-2+o,0! esmerilacao_!-3+mento,0! esnobismo_!-4+acao,0! esparramacao_!-4+e,0!
espetadela_!-3+a,0! espreitadela_!-3+a,0! esticadela_!-3+a,0! excitamento_!-5+cao,0! exoticidade_!-6+smo,0!
faiscancia_!-4+cao,0! faiscante_!0,-3+dor! falimento_!-6+encia,0! fatura_!+mento,0! fechada_!-2+mento,0!
fendimento_!-9+fissao,0! fereza_!-3+ocidade,0! finura_!-3+eza,0! fuzilacao_!-3+mento,0! fofice_!-3+ura,0!
forcamento_!-5,0! formatura_!-4+cao,0! fortitude_!-4+dao,0! fraturamento_!-5,0! frenacao_!-3+mento,0!
freamento_!-6+namento,0! frialdade_!-6+eza,0! frigidez_!-5+eza,0! furtadela_!-5+o,0! gravura_!-3+acao,0!
guiador_!0,-3! humanismo_!-3+dade,0! importunidade_!-4+ce,0! imprensada_!-1+ura,0! incensadela_!-4+cao,0!
incongruidade_!-5+encia,0! incrementacao_!-4+o,0! induzimento_!-7+cao,0! induzidor_!0,-5+tor! infiltramento_!-

5+cao,0! infinitude_!-4+dade,0! inquiricao_!-6+erito,0! interrogacao_!-3+torio,0! investida_!-2+mento,0! irrupcao_!-3+imento,0! laicismo_!-3+dade,0! lambidela_!-5+ecao,0! lambicao_!-4+ecao,0! lambuzadela_!-3+a,0! lancadura_!-4+mento,0! lance_!-1+amento,0! lavacao_!-3+gem,0! licenca_!-1+iamento,0! licenciatura_!-4+mento,0! ligeirismo_!-4+eza,0! limpadela_!-5+eza,0! limpamento_!-6+eza,0! lustradela_!-4+cao,0! madureza_!-6+turidade,0! maquilagem_!-5+agem,0! marca_!+cao,0! mascaragem_!-3+mento,0! mascara_!+mento,0! mexedura_!-4+cao,0! mineiridade_!-4+smo,0! misticidade_!-4+smo,0! mistica_!-1+ismo,0! misturacao_!-3,0! molhacao_!-3+mento,0! molhadela_!-4+mento,0! nivelacao_!-3+mento,0! nomeio_!-2+acao,0! orcada_!-2+mento,0! particularismo_!-3+dade,0! partimento_!-5+cao,0! passionalismo_!-3+dade,0! patenteacao_!-3+mento,0! penteadela_!-4+cao,0! percebimento_!-7+pcao,0! perdicao_!-4+a,0! picamento_!-5+da,0! pipoco_!-1+amento,0! pisadela_!-3+a,0! piscadela_!-3+a,0! polidura_!-4+mento,0! pouso_!-1+ada,0! prometimento_!-7+essa,0! proponente_!0,-5+sitor! proposta_!-2+icao,0! provacao_!-3,0! provadura_!-4,0! providencialidade_!-6,0! pungimento_!-7+cao,0! puxacao_!-3+da,0! puxamento_!-5+da,0! puxao_!-1+da,0! quebradela_!-4,0! quietacao_!-9+aquietacao,0! raladura_!-4+cao,0! ralada_!-2+cao,0! raspadela_!-3+a,0! rastejadura_!-4+mento,0! rebaix_!+amento,0! rebatimento_!-6+e,0! recalcamento_!-7+que,0! reduzida_!-4+cao,0! referimento_!-6+encia,0! refrescamento_!-6+o,0! refrigerio_!-4+acao,0! refusao_!-3+ndicao,0! regimento_!-6+encia,0! regedor_!0,-3+nte! regulagem_!-3+cao,0! relacionament_-8+ao,0! relaxacao_!-3+mento,0! relaxador_!0,-3+nte! remembramento_!-5+ca,0! removimento_!-6+ecao,0! rendicao_!-3+mento,0! renunciador_!0,-3+nte! replica_!+cao,0! reservacao_!-3,0! respancadura_!-4+mento,0! ressonadela_!-5+o,0! ressucitacao_!-3+mento,0! ressuncao_!-7+assuncao,0! resvaladura_!-4+mento,0! retiracao_!-3+da,0! ripadura_!-4+gem,0! ripamento_!-5+gem,0! roncadura_!-4+ria,0! rosnadela_!-3+a,0! ruptura_!-6+ompimento,0! sacad_!-3+que,0! sacacao_!-5+que,0! salpicadura_!-4+mento,0! salvacao_!-3+mento,0! sangradura_!-4+mento,0! sapiencia_!-7+bedoria,0! seladura_!-4+gem,0! selecionamento_!-9+ao,0! soldadura_!-4+gem,0! solta_!-1+ura,0! surramento_!-5,0! sustento_!-1+acao,0! tatura_!-3+eamento,0! tocadela_!-3+a,0! tombo_!-1+amento,0! torcidela_!-3+a,0! torra_!+cao,0! tosadura_!-4,0! tosquiadela_!-4,0! tossidela_!-5+e,0! tostadela_!3+ura,0! transa_!+cao,0! trilhamento_!-5,0! tropecao_!-1+mento,0! untadela_!-5+ura,0! validamento_!-5+cao,0! validade_!-2+cao,0! validez_!-2+acao,0! varred_!-2+ida,0! vasculhadela_!-3+a,0! visamento_!-5+da,0! visitacao_!-3,0! vista_!-2+ao,0! vitoria_!-6+encimento,0! vitorioso_!-8+encedor,0! volteadura_!-4+mento,0! voluntariado_!-3+edade,0! xingadela_!-4+cao,0!

ANEXO B REGRAS PARA IDENTIFICAÇÃO DE RLBS

São apresentadas, neste Anexo, as regras para identificação das RLBs, para o Português, utilizadas na ferramenta RELLEX.

Notação:

AA = adjetivo ou particípio

AJ = adjetivo

AP = particípio

AV = advérbio

CV = conjunto verbal

DT = determinante (artigos definido ou indefinido, ou pronomes demonstrativo ou indefinido)

LD = lado direito

LE = lado esquerdo

PR = preposição

SU = substantivo

VA = verbo auxiliar

VB = verbo

Regras para identificação de classificações

c1.Classificação direta:

$SU_1\ SU_2 \xrightarrow{rlb} = (\ SU_2, SU_1\ )$

Condição: há DT antes de SU_1, em LD ou LE, sem PR entre DT e SU_1.

Exemplo: o goleiro Manga $\xrightarrow{rlb}$ =(manga,goleiro)

c2. Classificação por verbo 'ser':

SU_1 'ser' $SU_2 \xrightarrow{rlb} = (\ SU_1, SU_2\ )$

Condição: não há núcleo no CV e SU_1 é núcleo no LE.

Exemplo: Manga foi goleiro $\xrightarrow{rlb}$ =(manga,goleiro)

c3. Classificação por predicado verbal:

SU VB $\xrightarrow{rlb} = (\ SU, \eta_2(VB)\ )$

Condição: SU é núcleo no LE, VB é núcleo no CV e não há preposição "por" no LD.

Exemplo: cidadão elegeu $\xrightarrow{rlb}$ =(cidadao,eleitor)

c4. Classificação por predicado nominal:

SU VA AA $\xrightarrow{rlb} = (\ SU, \eta_2(AA)\ )$

Condição: SU é núcleo no LE, AA é núcleo no CV e não há preposição "por" no LD.

Exemplo: animal é rastejante $\xrightarrow{rlb}$ =(animal,rastejador)

c5. Classificação do agente da voz passiva:

AP 'por' SU $\xrightarrow{rlb}$ = (SU, η_2(AP))

Condição: AP é núcleo no CV.

Exemplo: eleito pelo cidadão $\xrightarrow{rlb}$ =(cidadao,eleitor)

c6. Classificação por modificador

AA SU ou SU AA $\xrightarrow{rlb}$ de (AA, SU), se não há η_1(AA) nem η_2(AA)

Condição: mais próximo SU de AA, em LE ou LD, sem PR entre AA e SU.

Exemplo: biscoito crocante $\xrightarrow{rlb}$ =(biscoito,crocante)

Regras para identificação de restrições

r1. Restrição de objeto por modificador direto

AA SU ou SU AA $\xrightarrow{rlb}$ de (η_1(AA), SU), se há η_1(AA), ou
de (SU, η_2(AA)), se há η_2(AA)

Condição: mais próximo SU de AA, em LE ou LD, sem PR entre AA e SU.

Exemplos: equipe rápida $\xrightarrow{rlb}$ de(rapidez,equipe)
endereço residencial $\xrightarrow{rlb}$ de(endereco,residencia)

r2. Restrição de objeto por modificador preposicionado

SU_1 PR SU_2 $\xrightarrow{rlb}$ PR (SU_1, SU_2)

Condição: mais próximos SU_1 e SU_2 de PR, sem outra preposição antes deles.

Exemplo: fiscal com experiência $\xrightarrow{rlb}$ com(fiscal,experiência)

r3. Restrição de evento por modificador

AV VB ou VB AV $\xrightarrow{rlb}$ de (η_1(AV), η_1(VB)), se há η_1(AV), e
de (η_1(VB), η_2(AV)), se há η_2(AV)

Condição: VB é núcleo no CV.

Exemplos: projetou perfeitamente $\xrightarrow{rlb}$ de(perfeicao,projeto)
projetou mentalmente $\xrightarrow{rlb}$ de(projeto,mente)

r4. Restrição de modificador por modificador

AV AA ou AA AV $\xrightarrow{rlb}$ de (η_1(AV), η_1(AA)), se há η_1(AV), e
de (η_1(AA), η_2(AV)), se há η_2(AV)

Condição: AA é núcleo no CV ou, em LE ou LD, é o mais próximo AA de AV, sem PR entre eles.

Exemplos: adaptado rapidamente $\xrightarrow{rlb}$ de(rapidez,adaptacao)
adaptado pessoalmente $\xrightarrow{rlb}$ de(adaptacao,pessoa)

r5. Restrição de objeto por modificador de evento

SU VB AV $\xrightarrow{rlb}$ de (η_1(AV), SU), se há η_1(AV), senão
de (η_2(AV), SU), se há η_2(AV)

Condição: SU é núcleo em LE, VB é núcleo no CV e não há PR entre VB e AV.

Exemplos: atleta correu facilmente $\xrightarrow{rlb}$ de(facilidade,atleta)

feirante construiu artesanalmente $\xrightarrow{rlb}$ de(artesanato,feirante)

r6. Restrição de evento por agente

SU VB $\xrightarrow{rlb}$ por (η_1(VB), SU)

Condição: SU é núcleo no LE, VB é núcleo no CV e não há preposição "por" no LD.

Exemplo: forno esquentou $\xrightarrow{rlb}$ por(esquentamento,forno)

r7. Restrição de evento por tema

SU VB $\xrightarrow{rlb}$ de (η_1(VB), SU)

Condição: SU é núcleo no LE, VB é núcleo no CV e não há núcleo no LD.

Exemplo: forno esquentou $\xrightarrow{rlb}$ de(esquentamento,forno)

r8. Restrição de predicado nominal por agente

SU VA AA $\xrightarrow{rlb}$ por (η_1(AA), SU)

Condição: SU é núcleo no LE, AA é núcleo no CV e não há preposição "por" no LD.

Exemplo: prêmio tornou famoso $\xrightarrow{rlb}$ por(fama,premio)

r9. Restrição de predicado nominal por tema

SU VA AA $\xrightarrow{rlb}$ de (η_1(AA), SU)

Condição: SU é núcleo no LE, AA é núcleo no CV e não há núcleo no LD.

Exemplo: cantor ficou famoso $\xrightarrow{rlb}$ de(fama,cantor)

r10. Restrição de evento por objeto

VB SU $\xrightarrow{rlb}$ de (η_1(VB), SU)

Condição: SU é núcleo no LD, VB é núcleo no CV.

Exemplo: comprei presente $\xrightarrow{rlb}$ de(compra,presente)

r11. Restrição de predicado nominal por objeto

VA AA SU $\xrightarrow{rlb}$ de (η_1(AA), SU)

Condição: SU é núcleo no LD, AA é núcleo no CV.

Exemplo: foi comprado o presente $\xrightarrow{rlb}$ de(compra,presente)

r12. Restrição de evento por complemento

VB PR SU $\xrightarrow{rlb}$ PR (η_1(VB), SU)

Condição: VB é núcleo no CV, PR é primeira preposição no LD e SU é o primeiro substantivo após PR.

Exemplo: comprei na loja $\xrightarrow{rlb}$ em(compra,loja)

r13. Restrição de predicado nominal por complemento

VA AA PR SU $\xrightarrow{rlb}$ PR (η_1(AA), SU)

Condição: AA é núcleo no CV, PR é primeira preposição no LD e SU é o primeiro substantivo após PR.

Exemplo: ficou calmo sobre a cama $\xrightarrow{rlb}$ sobre(calma,cama)

r14. Restrição de agente por complemento

SU_1 VA PR SU_2 $\xrightarrow{rlb}$ PR (SU1, SU_2)

Condição: SU_1 é núcleo no LE, não há núcleo no CV, PR é primeira preposição no LD e SU_2 é o primeiro substantivo após PR.

Exemplo: equipe está na competição $\xrightarrow{rlb}$ em(equipe,competição)

r15. Restrição de possuído por possuidor

SU_1 'ter/possuir' SU_2 $\xrightarrow{rlb}$ de (SU_2, SU_1)

Condição: SU_1 é núcleo no LE e SU2 é núcleo no LD.

Exemplo: casa tem porta $\xrightarrow{rlb}$ de(porta,casa)

Regras para identificação de associações

a1. Associação de agente com tema em evento

SU_1 VB SU_2 $\xrightarrow{rlb}$ η_1(VB) (SU1, SU_2)

Condição: SU_1 é núcleo no LE, VB é núcleo no CV e SU_2 é núcleo no LD.

Exemplo: técnico treinou atleta $\xrightarrow{rlb}$ treino(tecnico,atleta)

a2. Associação de agente com tema na voz passiva

SU_1 VA AA 'por' SU_2 $\xrightarrow{rlb}$ η_1(AA) (SU1, SU_2)

Condição: SU_1 é núcleo no LE, AA é núcleo no CV e SU_2 é núcleo no LD.

Exemplo: atleta foi treinado pelo técnico $\xrightarrow{rlb}$ treino(tecnico,atleta)

a3. Associação de agente com tema em evento preposicionado

SU_1 VB PR SU_2 $\xrightarrow{rlb}$ η_1(VB).PR (SU1, SU_2)

Condição: SU_1 é núcleo no LE, VB é núcleo no CV, PR é primeira preposição no LD e SU_2 é o primeiro substantivo após PR.

Exemplo: turista viajou para a Europa $\xrightarrow{rlb}$ viagem.para(turista,europa)

ANEXO C TRECHOS DE ARQUIVOS DE ÍNDICE

É apresentado, neste Anexo, um trecho de um documento da coleção utilizada em formato normal e na forma verticalizada. Também são apresentados trechos de arquivos de índice correspondentes a este texto.

Trecho do documento 1410:

```
...
O boneco de madeira Pinóquio, criado por Gepetto, está no teatro.
...
```

Trecho verticalizado do documento 1410 no conjunto de documentos da coleção de referência Folha94:

```
...
O o o 0 0 _AD
boneco boneco bonec 0 0 _SU
de de = 0 0 _PR
madeira madeira madeir 0 0 _SU
Pinóquio pinoquio pinoq 0 0 _SU
, , = 0 0 _VG
criado criar cri criacao criador _AP
por por = 0 0 _PR
Gepetto gepetto gepett 0 0 _SU
, , = 0 0 _VG
está estar est 0 0 _VA
em em = 0 0 _PR
o o o 0 0 _AD
teatro teatro teatr 0 0 _SU
. . = 0 0 _PN
...
```

Notação (para o texto verticalizado):

= significa *stem* igual à palavra original,

0 significa ausência de nominalização e

_XX significa etiqueta (conforme Tabela 4.2).

A ordem com que as informações são apresentadas é a seguinte: palavra, lema, *stem*, substantivo abstrato, substantivo concreto e etiqueta morfológica.

Trechos dos arquivos de índice de cada estratégia de indexação:

São apresentados, a seguir, trechos onde descritores do documento 1410 aparecem. Dados referentes ao documento 1410 estão em negrito.

Estratégia de indexação VR:

```
(conteúdo: descritor IDF documento freqüência ...)
boneco 6.03 3967 1 1926 1 1439 1 1421 1 1418 3 1410 2 1185 1 933 1 310 2
   304 1
```

```
criado 3.93 4152 1 3966 1 3946 1 3904 1 3888 1 3808 1 3562 1 3516 1 3374
   1 3361 1 3342 1 3289 1 3282 1 3279 1 3246 1 3221 1 3219 1 3156 1 3141
   1 3024 1 3009 1 2827 1 2810 1 2763 1 2693 1 2583 1 2581 1 2572 1 2436
   1 2421 1 2412 1 2212 1 2158 1 2143 1 2125 1 2080 1 2049 1 2015 1 2011
   1 1989 1 1981 1 1945 1 1925 1 1924 1 1904 1 1878 1 1869 1 1819 1 1816
   1 1777 1 1685 1 1530 1 1528 1 1510 1 1508 1 1464 1 1416 1 1410 1 1353
   1 1251 1 1204 1 942 1 936 1 854 1 844 1 798 2 797 1 796 1 745 1 684 2
   654 1 645 1 636 1 628 1 429 2 282 1 266 1 169 1 108 2 82 1 75 1 65 1
gepetto 8.33 1410 2
madeira 4.25 4137 1 4135 1 4044 1 4030 1 3980 1 3927 1 3894 1 3868 1 3867
   1 3863 2 3818 1 3802 1 3756 1 3753 3 3711 1 3690 2 3689 1 3687 1 3654
   3 3598 4 3596 1 3587 7 3442 1 3329 1 3319 1 3307 1 3276 2 3265 1 3199
   1 3191 1 3178 1 3173 1 3137 1 3089 2 2902 1 2881 1 2807 1 2780 1 2630
   1 2564 1 2406 1 2301 1 1885 1 1844 1 1814 2 1813 1 1812 4 1807 2 1802
   3 1630 1 1562 1 1559 1 1429 2 1410 1 949 1 492 1 465 1 238 2 137 2
pinoquio 7.23 2943 1 1510 1 1410 3
teatro 3.67 4034 2 4031 1 3954 2 3950 1 3904 2 3856 1 3774 1 3749 1 3616
   1 3545 2 3544 1 3526 1 3521 1 3504 1 3496 1 3400 1 3388 1 3322 3 3305
   1 3298 1 3288 6 3254 6 3232 4 3217 1 3213 2 3212 1 3187 1 3184 2 3164
   2 3162 1 3146 3 3145 2 3117 1 3112 1 3108 1 3101 1 3093 1 3073 1 3064
   1 3046 1 3041 1 3010 1 2999 1 2907 1 2682 2 2680 1 2652 9 2651 7 2647
   2 2621 1 2618 1 2561 3 2530 1 2485 5 2481 1 2477 3 2465 1 2440 1 2385
   1 2328 1 2325 1 2322 1 2317 4 2252 1 2251 4 2250 6 2153 1 2070 1 1786
   1 1755 8 1740 1 1734 1 1727 6 1710 1 1690 5 1688 1 1683 2 1681 1 1680
   3 1676 2 1648 1 1614 1 1605 1 1601 2 1593 1 1590 2 1587 3 1578 1 1577
   1 1552 1 1538 1 1534 1 1507 1 1479 1 1474 1 1467 1 1464 1 1414 4 1410
   2 1409 1 1408 1 1407 2 1101 1 872 1 842 1 339 1
```

Estratégia de indexação LM:

```
(conteúdo: descritor IDF documento freqüência ...)
boneco 5.39 3980 3 3979 1 3967 2 3958 6 3886 1 3761 2 2273 2 1926 2 1439
   1 1426 1 1421 1 1418 6 1410 2 1407 2 1185 1 933 1 357 4 310 4 304 1
criar 2.25 4152 1 4105 1 4090 1 4030 1 3966 1 3964 1 3959 1 3946 1 3936 1
   3926 1 3904 1 3894 1 3888 1 3874 1 3851 1 3843 5 3825 1 3815 1 3808 1
   3772 1 3764 1 3761 1 3757 1 3745 1 3733 1 3719 1 3718 1 3678 1 3669 1
   3596 1 3576 2 3562 1 3542 1 3529 1 3521 1 3518 1 3516 1 3510 1 3497 2
   3495 1 3490 1 3473 2 3445 1 3442 1 3436 2 3419 1 3405 1 3403 1 3401 1
   3384 2 3379 1 3374 2 3361 1 3347 3 3344 1 3342 1 3329 1 3312 1 3289 1
   3288 1 3282 1 3279 2 3277 1 3272 1 3264 1 3246 1 3239 1 3225 1 3221 1
   3219 1 3216 1 3213 1 3197 1 3192 1 3191 1 3190 2 3183 1 3178 1 3156 1
   3145 1 3141 1 3134 1 3132 2 3130 1 3122 1 3081 1 3079 1 3078 2 3076 1
   3060 1 3056 1 3024 1 3016 1 3013 1 3010 1 2990 1 2964 1 2954 1 2937 1
   2899 1 2891 1 2849 1 2844 1 2827 1 2826 1 2823 1 2818 2 2816 2 2810 3
   2805 1 2795 1 2787 1 2785 1 2763 1 2761 1 2760 1 2743 2 2727 1 2721 1
   2710 1 2705 1 2702 1 2696 1 2693 1 2691 2 2678 3 2676 2 2663 1 2653 1
   2647 1 2640 1 2635 1 2622 1 2588 1 2587 1 2586 1 2583 1 2581 3 2572 1
   2564 1 2552 2 2551 1 2548 1 2546 2 2545 1 2538 2 2530 1 2529 3 2528 1
   2527 1 2521 1 2519 1 2518 1 2504 2 2496 1 2491 1 2476 2 2474 1 2468 2
   2467 1 2452 1 2449 2 2443 1 2436 1 2434 2 2433 1 2432 1 2431 1 2421 1
   2412 1 2409 1 2404 1 2401 1 2399 1 2397 1 2388 1 2385 1 2379 2 2367 1
   2357 2 2351 1 2350 3 2349 1 2348 1 2345 1 2331 1 2327 1 2326 1 2323 1
   2321 1 2317 1 2305 1 2301 1 2299 1 2298 1 2275 2 2268 1 2262 1 2260 1
   2249 4 2233 1 2220 1 2217 1 2212 2 2211 1 2208 1 2196 2 2187 1 2185 1
   2181 1 2173 1 2172 3 2171 1 2170 1 2161 2 2159 1 2158 1 2143 1 2133 1
   2127 1 2126 1 2125 1 2085 1 2083 2 2080 2 2068 1 2065 2 2049 1 2044 1
   2015 2 2011 3 2010 4 2009 1 2001 2 1995 1 1989 1 1986 1 1983 2 1981 1
   1975 1 1957 1 1955 2 1950 2 1945 3 1944 2 1942 1 1937 1 1934 1 1928 1
   1927 1 1926 1 1925 1 1924 1 1913 1 1904 1 1903 1 1885 1 1880 1 1878 2
   1869 1 1842 1 1841 1 1839 1 1835 1 1831 1 1821 1 1819 1 1816 1 1814 1
   1809 1 1795 1 1789 1 1777 1 1767 1 1754 1 1747 1 1746 1 1744 1 1737 1
   1727 3 1725 1 1724 1 1719 1 1695 1 1693 1 1686 1 1685 1 1653 1 1652 1
   1646 1 1640 2 1619 1 1614 1 1606 3 1605 1 1602 1 1593 1 1590 1 1578 3
   1577 3 1572 1 1558 2 1557 1 1552 1 1533 1 1530 2 1528 1 1517 1 1510 1
   1508 1 1507 1 1481 2 1472 1 1464 1 1439 1 1437 1 1433 1 1418 1 1416 1
   1410 1 1399 1 1387 1 1374 1 1353 1 1348 1 1317 1 1257 1 1251 1 1249 3
```

```
   1245 1 1221 1 1219 1 1204 1 1166 1 1157 1 1116 1 1089 1 1070 1 1051 1
   974 1 970 1 968 1 949 1 942 1 936 1 933 1 915 1 912 1 908 1 867 1 862
   1 858 1 854 1 853 1 851 2 848 2 844 2 837 1 830 1 829 1 819 1 798 4
   797 1 796 1 791 1 787 2 785 3 784 3 760 1 750 2 745 1 733 1 715 2 700
   1 699 1 690 1 684 2 667 1 654 2 645 1 636 3 632 1 630 1 628 1 621 1
   614 2 581 2 576 1 571 1 569 3 568 1 527 3 513 1 473 1 469 1 462 1 429
   2 422 1 330 1 329 1 322 1 321 1 317 1 314 1 282 1 277 2 266 1 209 1
   207 1 203 1 199 2 186 1 184 1 170 2 169 1 147 1 123 1 119 1 116 1 109
   2 108 2 94 1 84 1 83 1 82 1 77 1 69 1 67 2 65 1 63 2 54 1 46 1 39 1 29
   1 12 1 7 1 6 1
gepetto 8.33 1410 2
madeira 4.22 4137 1 4135 1 4044 1 4030 1 3927 1 3894 1 3868 1 3867 1 3863
   2 3818 1 3802 1 3756 1 3753 3 3711 1 3690 2 3689 1 3687 1 3654 3 3598
   4 3596 1 3587 7 3442 1 3329 1 3319 1 3307 1 3277 1 3276 2 3265 1 3199
   1 3191 1 3178 1 3173 1 3137 1 3089 2 2902 2 2881 1 2807 1 2780 1 2630
   1 2564 1 2406 1 2301 1 1885 1 1844 1 1814 3 1813 1 1812 5 1807 2 1802
   3 1630 1 1562 1 1559 1 1558 2 1429 2 1415 1 1410 1 949 1 492 1 465 1
   238 2 137 2
pinoquio 7.23 2943 1 1510 1 1410 3
teatro 3.62 4034 2 4031 1 3954 2 3950 1 3926 1 3904 2 3903 1 3856 1 3774
   1 3749 1 3616 1 3545 2 3544 1 3526 1 3521 1 3504 1 3496 1 3495 1 3400
   1 3388 1 3322 3 3305 1 3298 1 3288 6 3254 6 3232 4 3217 1 3213 2 3212
   1 3187 1 3184 2 3164 2 3162 1 3146 3 3145 2 3117 1 3112 1 3108 1 3101
   1 3093 1 3073 1 3064 1 3046 1 3041 1 3015 1 3010 1 2999 1 2907 1 2769
   1 2682 2 2680 1 2652 9 2651 7 2647 2 2621 1 2618 1 2561 3 2530 1 2485
   5 2481 1 2477 3 2465 1 2440 1 2385 1 2328 1 2325 1 2322 1 2317 4 2252
   1 2251 4 2250 6 2153 1 2070 1 1786 1 1755 8 1740 1 1734 1 1727 6 1710
   1 1690 5 1688 1 1683 2 1681 1 1680 3 1676 2 1648 1 1614 1 1605 1 1601
   2 1593 1 1590 2 1587 3 1578 1 1577 1 1552 1 1538 1 1534 1 1507 1 1479
   1 1474 1 1467 1 1464 1 1414 4 1410 2 1409 1 1408 1 1407 2 1101 1 872 1
   842 1 339 1
```

Estratégia de indexação ST:

```
(conteúdo: descritor IDF documento freqüência ...)
bonec 5.07 3980 3 3979 1 3967 3 3958 6 3886 1 3761 3 3627 1 3329 1 3194 1
   2277 1 2273 2 1926 5 1745 1 1439 1 1426 1 1425 1 1421 2 1418 9 1410 2
   1408 1 1407 2 1185 1 933 1 357 4 310 4 304 1
cri 1.91 4152 1 4128 1 4105 1 4090 1 4045 1 4040 1 4030 1 4011 1 3997 1
   3966 1 3964 1 3959 1 3957 1 3946 1 3936 1 3926 1 3919 1 3904 1 3895 1
   3894 1 3888 1 3885 1 3874 1 3861 1 3851 1 3843 5 3825 1 3815 1 3808 1
   3772 1 3766 1 3765 1 3764 1 3761 1 3757 1 3745 1 3733 1 3730 1 3719 1
   3718 1 3679 1 3678 1 3669 1 3646 1 3596 1 3576 2 3565 1 3562 1 3542 1
   3540 2 3539 2 3529 1 3521 1 3518 2 3516 1 3510 1 3507 3 3497 2 3495 1
   3490 1 3473 2 3445 2 3442 1 3436 2 3419 1 3405 1 3403 1 3401 1 3384 2
   3379 1 3374 4 3371 1 3361 1 3349 1 3347 3 3344 1 3342 1 3329 1 3312 1
   3305 1 3289 1 3288 1 3285 1 3282 1 3279 2 3277 1 3272 1 3266 2 3264 1
   3246 1 3239 1 3225 1 3221 2 3219 2 3216 1 3213 1 3197 1 3192 1 3191 1
   3190 2 3183 1 3181 1 3178 1 3175 1 3156 2 3145 1 3141 1 3134 1 3132 2
   3131 1 3130 1 3122 1 3082 1 3081 2 3079 1 3078 2 3076 1 3060 2 3056 2
   3040 1 3039 3 3034 1 3024 1 3016 1 3013 1 3010 1 3009 1 3004 1 2994 1
   2990 1 2988 1 2964 1 2954 1 2937 1 2910 5 2899 1 2898 1 2891 1 2849 1
   2844 1 2840 1 2827 1 2826 1 2823 1 2818 4 2817 2 2816 2 2813 1 2810 3
   2805 1 2795 1 2792 1 2789 1 2787 2 2785 1 2769 1 2763 1 2761 1 2760 1
   2743 2 2730 1 2727 1 2721 2 2710 1 2705 1 2702 1 2696 2 2693 1 2691 2
   2684 1 2680 1 2678 3 2676 2 2675 1 2663 1 2653 2 2652 1 2651 1 2647 2
   2646 1 2643 1 2640 1 2637 1 2635 3 2622 1 2619 1 2614 1 2603 1 2594 1
   2588 2 2587 3 2586 1 2583 2 2581 3 2572 1 2564 8 2561 1 2556 1 2555 1
   2552 2 2551 1 2548 1 2547 2 2546 2 2545 1 2540 1 2538 2 2532 1 2530 2
   2529 6 2528 1 2527 1 2521 1 2519 3 2518 1 2517 1 2510 1 2504 2 2496 1
   2493 1 2492 1 2491 1 2485 1 2479 2 2477 1 2476 2 2474 1 2468 2 2467 1
   2464 1 2454 1 2452 2 2450 1 2449 2 2443 3 2438 1 2436 1 2434 2 2433 1
   2432 1 2431 1 2421 2 2415 1 2412 3 2409 1 2404 1 2401 1 2399 2 2397 2
   2388 1 2385 1 2379 2 2367 1 2358 1 2357 2 2351 4 2350 4 2349 1 2348 1
   2345 1 2337 1 2331 1 2330 1 2327 2 2326 1 2323 2 2321 1 2317 1 2305 1
   2303 1 2302 1 2301 2 2299 1 2298 1 2284 1 2275 2 2268 1 2265 1 2262 1
```

```
   2260 1 2253 1 2251 2 2249 5 2247 1 2246 2 2236 1 2235 1 2233 1 2220 1
   2218 1 2217 1 2212 2 2211 1 2208 2 2196 2 2187 1 2185 1 2181 1 2180 1
   2174 1 2173 1 2172 3 2171 2 2170 1 2161 3 2159 1 2158 1 2143 1 2134 1
   2133 1 2127 1 2126 1 2125 1 2113 1 2092 2 2090 1 2085 1 2083 3 2080 2
   2079 1 2068 1 2065 4 2049 2 2044 1 2015 2 2011 3 2010 4 2009 1 2001 2
   1995 1 1989 1 1986 1 1983 2 1982 1 1981 1 1975 1 1963 1 1957 1 1955 2
   1950 3 1945 4 1944 2 1942 1 1937 1 1934 1 1928 2 1927 1 1926 1 1925 1
   1924 2 1918 1 1913 1 1904 1 1903 1 1885 1 1881 1 1880 3 1878 2 1869 1
   1842 1 1841 1 1839 1 1835 1 1831 1 1821 1 1819 1 1816 1 1814 1 1809 1
   1795 1 1789 1 1777 2 1767 1 1765 1 1754 1 1747 1 1746 1 1744 1 1743 4
   1738 1 1737 1 1734 1 1727 3 1725 2 1724 1 1723 1 1719 1 1708 1 1695 1
   1693 1 1686 1 1685 1 1675 1 1671 1 1661 1 1653 1 1652 1 1646 1 1643 1
   1641 1 1640 2 1619 1 1614 1 1606 3 1605 1 1602 1 1593 3 1590 1 1586 1
   1584 1 1578 5 1577 5 1572 2 1565 1 1558 3 1557 1 1552 1 1550 1 1542 1
   1536 1 1533 1 1530 2 1528 1 1517 1 1510 1 1508 1 1507 1 1494 1 1481 2
   1472 1 1464 1 1462 1 1439 1 1437 1 1433 1 1418 1 1416 1 1410 1 1400 2
   1399 1 1387 1 1374 1 1353 1 1348 1 1324 1 1317 1 1257 1 1251 1 1249 3
   1245 1 1241 1 1221 1 1219 1 1204 1 1186 1 1166 1 1157 1 1155 1 1116 1
   1089 1 1070 1 1051 1 974 1 970 1 968 2 949 1 942 1 936 1 933 1 928 1
   915 1 914 1 912 1 908 1 889 1 867 1 862 1 858 2 856 1 854 1 853 1 851
   2 848 2 844 2 839 1 837 1 830 1 829 2 823 1 819 1 815 2 798 11 797 1
   796 1 795 1 792 1 791 1 787 2 785 4 784 3 767 1 760 1 750 2 745 1 733
   1 715 2 700 1 699 1 690 1 687 1 684 2 677 1 667 1 654 2 645 2 636 3
   632 2 630 1 628 1 621 1 614 2 608 1 603 1 581 2 578 3 577 2 576 1 571
   1 569 6 568 2 548 1 527 6 513 1 473 1 469 1 462 2 446 1 429 2 422 1
   413 1 330 1 329 1 324 1 322 1 321 1 317 1 314 1 292 1 291 2 282 1 277
   2 266 1 247 1 209 1 207 1 203 1 199 2 186 1 184 1 170 2 169 1 147 3
   123 1 122 1 119 2 116 1 114 3 113 1 112 2 111 3 109 6 108 4 106 2 105
   1 96 1 94 1 91 1 89 1 84 4 83 4 82 1 81 3 79 1 78 1 77 4 76 1 75 11 74
   4 69 2 68 1 67 5 66 2 65 4 64 3 63 3 62 1 60 1 59 4 58 2 54 1 46 1 39
   3 29 1 28 5 27 1 25 1 24 1 23 1 12 1 10 1 8 1 7 1 6 2
gepett 8.33 1410 2
madeir 4.21 4137 1 4135 1 4044 1 4030 1 3980 1 3927 1 3894 1 3868 1 3867
   1 3863 2 3818 1 3802 1 3756 1 3753 3 3711 1 3690 2 3689 1 3687 1 3654
   3 3598 4 3596 1 3587 7 3442 1 3329 1 3319 1 3307 1 3277 1 3276 2 3265
   1 3199 1 3191 1 3178 1 3173 1 3137 1 3089 2 2902 2 2881 1 2807 1 2780
   1 2630 1 2564 1 2406 1 2301 1 1885 1 1844 1 1814 3 1813 1 1812 5 1807
   2 1802 3 1630 1 1562 1 1559 1 1558 2 1429 2 1415 1 1410 1 949 1 492 1
   465 1 238 6 137 2
pinoq 7.23 2943 1 1510 1 1410 3
teatr 3.53 4037 1 4034 3 4031 1 4018 1 3954 2 3950 1 3926 1 3904 2 3903 1
   3856 1 3774 1 3749 1 3688 1 3616 1 3545 2 3544 1 3526 1 3521 1 3504 1
   3496 1 3495 1 3400 1 3388 1 3322 3 3305 1 3298 1 3288 6 3266 1 3254 7
   3232 4 3217 1 3215 1 3213 2 3212 1 3187 1 3184 2 3164 2 3162 1 3146 3
   3145 2 3117 1 3112 1 3108 1 3101 1 3093 1 3073 1 3064 1 3062 2 3046 1
   3041 1 3015 1 3010 1 2999 1 2907 2 2769 1 2682 2 2680 1 2652 9 2651 9
   2647 2 2621 1 2618 1 2561 5 2530 1 2485 5 2481 1 2477 4 2465 1 2440 1
   2412 1 2385 1 2328 1 2325 1 2322 1 2317 5 2312 1 2265 1 2252 1 2251 4
   2250 8 2153 1 2070 1 1786 1 1755 9 1740 1 1734 1 1727 6 1710 4 1690 5
   1688 1 1683 2 1681 1 1680 3 1676 2 1648 1 1614 1 1605 1 1601 2 1593 1
   1590 2 1587 3 1578 1 1577 1 1552 1 1538 1 1534 1 1507 1 1479 1 1474 1
   1467 1 1464 1 1414 4 1410 2 1409 1 1408 1 1407 2 1101 1 1031 1 872 1
   842 1 741 2 339 1
```

Estratégia de indexação NM:

```
(conteúdo: descritor IDF documento freqüência ...)
boneco 5.39 3980 3 3979 1 3967 2 3958 6 3886 1 3761 2 2273 2 1926 2 1439
   1 1426 1 1421 1 1418 6 1410 2 1407 2 1185 1 933 1 357 4 310 4 304 1
criacao 2.04 4152 1 4128 1 4105 1 4090 1 4045 1 4030 1 3997 1 3966 1 3964
   1 3959 1 3946 1 3936 1 3926 1 3904 1 3895 1 3894 1 3888 1 3885 1 3874
   1 3861 1 3851 1 3843 5 3825 1 3815 1 3808 1 3772 1 3766 1 3764 1 3761
   1 3757 1 3745 1 3733 1 3719 1 3718 1 3678 1 3669 1 3596 1 3576 2 3565
   1 3562 1 3542 1 3540 2 3539 2 3529 1 3521 1 3518 1 3516 1 3510 1 3497
   2 3495 1 3490 1 3473 2 3445 2 3442 1 3436 2 3419 1 3405 1 3403 1 3401
   1 3384 2 3379 1 3374 2 3361 1 3349 1 3347 3 3344 1 3342 1 3329 1 3312
```

1 3305 1 3289 1 3288 1 3282 1 3279 2 3277 1 3272 1 3266 2 3264 1 3246
1 3239 1 3225 1 3221 2 3219 2 3216 1 3213 1 3197 1 3192 1 3191 1 3190
2 3183 1 3178 1 3156 1 3145 1 3141 1 3134 1 3132 2 3130 1 3122 1 3082
1 3081 2 3079 1 3078 2 3076 1 3060 2 3056 1 3040 1 3039 2 3034 1 3024
1 3016 1 3013 1 3010 1 2990 1 2988 1 2964 1 2954 1 2937 1 2899 1 2898
1 2891 1 2849 1 2844 1 2840 1 2827 1 2826 1 2823 1 2818 2 2816 2 2813
1 2810 3 2805 1 2795 1 2792 1 2789 1 2787 2 2785 1 2763 1 2761 1 2760
1 2743 2 2730 1 2727 1 2721 2 2710 1 2705 1 2702 1 2696 1 2693 1 2691
2 2684 1 2678 3 2676 2 2675 1 2663 1 2653 2 2647 2 2646 1 2643 1 2640
1 2635 1 2622 1 2619 1 2614 1 2603 1 2594 1 2588 2 2587 3 2586 1 2583
2 2581 3 2572 1 2564 6 2561 1 2556 1 2552 2 2551 1 2548 1 2547 1 2546
2 2545 1 2538 2 2532 1 2530 2 2529 6 2528 1 2527 1 2521 1 2519 2 2518
1 2517 1 2504 2 2496 1 2493 1 2492 1 2491 1 2485 1 2479 2 2477 1 2476
2 2474 1 2468 2 2467 1 2464 1 2454 1 2452 2 2450 1 2449 2 2443 2 2438
1 2436 1 2434 2 2433 1 2432 1 2431 1 2421 2 2415 1 2412 1 2409 1 2404
1 2401 1 2399 1 2397 1 2388 1 2385 1 2379 2 2367 1 2357 2 2351 4 2350
4 2349 1 2348 1 2345 1 2337 1 2331 1 2330 1 2327 1 2326 1 2323 2 2321
1 2317 1 2305 1 2302 1 2301 1 2299 1 2298 1 2284 1 2275 2 2268 1 2265
1 2262 1 2260 1 2253 1 2249 5 2247 1 2246 1 2233 1 2220 1 2217 1 2212
2 2211 1 2208 2 2196 2 2187 1 2185 1 2181 1 2180 1 2174 1 2173 1 2172
3 2171 2 2170 1 2161 3 2159 1 2158 1 2143 1 2134 1 2133 1 2127 1 2126
1 2125 1 2092 2 2090 1 2085 1 2083 3 2080 2 2079 1 2068 1 2065 3 2049
2 2044 1 2015 2 2011 3 2010 4 2009 1 2001 2 1995 1 1989 1 1986 1 1983
2 1981 1 1975 1 1957 1 1955 2 1950 3 1945 3 1944 2 1942 1 1937 1 1934
1 1928 2 1927 1 1926 1 1925 1 1924 1 1918 1 1913 1 1904 1 1903 1 1885
1 1881 1 1880 3 1878 2 1869 1 1842 1 1841 1 1839 1 1835 1 1831 1 1821
1 1819 1 1816 1 1814 1 1809 1 1795 1 1789 1 1777 2 1767 1 1765 1 1754
1 1747 1 1746 1 1744 1 1743 3 1738 1 1737 1 1734 1 1727 3 1725 1 1724
1 1719 1 1695 1 1693 1 1686 1 1685 1 1675 1 1671 1 1661 1 1653 1 1652
1 1646 1 1640 2 1619 1 1614 1 1606 3 1605 1 1602 1 1593 2 1590 1 1578
4 1577 4 1572 2 1565 1 1558 2 1557 1 1552 1 1550 1 1542 1 1536 1 1533
1 1530 2 1528 1 1517 1 1510 1 1508 1 1507 1 1481 2 1472 1 1464 1 1439
1 1437 1 1433 1 1418 1 1416 1 **1410 1** 1399 1 1387 1 1374 1 1353 1 1348
1 1324 1 1317 1 1257 1 1251 1 1249 3 1245 1 1241 1 1221 1 1219 1 1204
1 1166 1 1157 1 1155 1 1116 1 1089 1 1070 1 1051 1 974 1 970 1 968 1
949 1 942 1 936 1 933 1 928 1 915 1 914 1 912 1 908 1 889 1 867 1 862
1 858 2 854 1 853 1 851 2 848 2 844 2 839 1 837 1 830 1 829 1 823 1
819 1 798 11 797 1 796 1 795 1 792 1 791 1 787 2 785 3 784 3 767 1 760
1 750 2 745 1 733 1 715 2 700 1 699 1 690 1 687 1 684 2 677 1 667 1
654 2 645 2 636 3 632 1 630 1 628 1 621 1 614 2 608 1 603 1 581 2 578
3 577 2 576 1 571 1 569 6 568 2 548 1 527 3 513 1 473 1 469 1 462 2
446 1 429 2 422 1 413 1 330 1 329 1 324 1 322 1 321 1 317 1 314 1 292
1 291 2 282 1 277 2 266 1 247 1 209 1 207 1 203 1 199 2 186 1 184 1
170 2 169 1 147 2 123 1 119 2 116 1 114 1 109 3 108 2 94 1 84 3 83 1
82 1 81 1 77 2 75 2 69 1 67 2 66 1 65 1 63 2 54 1 46 1 39 1 29 1 12 1
8 1 7 1 6 2
criador 2.15 4152 1 4105 1 4090 1 4040 1 4030 1 4011 1 3966 1 3964 1 3959
1 3957 1 3946 1 3936 1 3926 1 3904 1 3894 1 3888 1 3874 1 3851 1 3843
5 3825 1 3815 1 3808 1 3772 1 3765 1 3764 1 3761 1 3757 1 3745 1 3733
1 3719 1 3718 1 3679 1 3678 1 3669 1 3596 1 3576 2 3562 1 3542 1 3529
1 3521 1 3518 2 3516 1 3510 1 3497 2 3495 1 3490 1 3473 2 3445 1 3442
1 3436 2 3419 1 3405 1 3403 1 3401 1 3384 2 3379 1 3374 4 3361 1 3347
3 3344 1 3342 1 3329 1 3312 1 3289 1 3288 1 3282 1 3279 2 3277 1 3272
1 3264 1 3246 1 3239 1 3225 1 3221 1 3219 1 3216 1 3213 1 3197 1 3192
1 3191 1 3190 2 3183 1 3178 1 3156 1 3145 1 3141 1 3134 1 3132 2 3130
1 3122 1 3081 1 3079 1 3078 2 3076 1 3060 1 3056 1 3039 1 3024 1 3016
1 3013 1 3010 1 3004 1 2990 1 2964 1 2954 1 2937 1 2899 1 2891 1 2849
1 2844 1 2827 1 2826 1 2823 1 2818 2 2816 2 2810 3 2805 1 2795 1 2787
1 2785 1 2763 1 2761 1 2760 1 2743 2 2727 1 2721 1 2710 1 2705 1 2702
1 2696 1 2693 1 2691 2 2680 1 2678 3 2676 2 2663 1 2653 1 2647 1 2640
1 2635 1 2622 1 2588 1 2587 1 2586 1 2583 1 2581 3 2572 1 2564 2 2552
2 2551 1 2548 1 2546 2 2545 1 2538 2 2530 1 2529 3 2528 1 2527 1 2521
1 2519 1 2518 1 2504 2 2496 1 2491 1 2476 2 2474 1 2468 2 2467 1 2452
1 2449 2 2443 2 2436 1 2434 2 2433 1 2432 1 2431 1 2421 1 2412 1 2409
1 2404 1 2401 1 2399 1 2397 1 2388 1 2385 1 2379 2 2367 1 2357 2 2351
1 2350 3 2349 1 2348 1 2345 1 2331 1 2327 2 2326 1 2323 1 2321 1 2317
1 2305 1 2301 2 2299 1 2298 1 2275 2 2268 1 2262 1 2260 1 2249 4 2236
1 2233 1 2220 1 2217 1 2212 2 2211 1 2208 1 2196 2 2187 1 2185 1 2181

```
   1 2173 1 2172 3 2171 1 2170 1 2161 2 2159 1 2158 1 2143 1 2133 1 2127
   1 2126 1 2125 1 2113 1 2085 1 2083 2 2080 2 2068 1 2065 2 2049 1 2044
   1 2015 2 2011 3 2010 4 2009 1 2001 2 1995 1 1989 1 1986 1 1983 2 1982
   1 1981 1 1975 1 1963 1 1957 1 1955 2 1950 2 1945 4 1944 2 1942 1 1937
   1 1934 1 1928 1 1927 1 1926 1 1925 1 1924 2 1913 1 1904 1 1903 1 1885
   1 1880 1 1878 2 1869 1 1842 1 1841 1 1839 1 1835 1 1831 1 1821 1 1819
   1 1816 1 1814 1 1809 1 1795 1 1789 1 1777 1 1767 1 1754 1 1747 1 1746
   1 1744 1 1743 1 1737 1 1727 3 1725 1 1724 1 1719 1 1708 1 1695 1 1693
   1 1686 1 1685 1 1653 1 1652 1 1646 1 1643 1 1640 2 1619 1 1614 1 1606
   3 1605 1 1602 1 1593 2 1590 1 1584 1 1578 4 1577 4 1572 1 1558 3 1557
   1 1552 1 1533 1 1530 2 1528 1 1517 1 1510 1 1508 1 1507 1 1481 2 1472
   1 1464 1 1439 1 1437 1 1433 1 1418 1 1416 1 1410 1 1399 1 1387 1 1374
   1 1353 1 1348 1 1317 1 1257 1 1251 1 1249 3 1245 1 1221 1 1219 1 1204
   1 1166 1 1157 1 1116 1 1089 1 1070 1 1051 1 974 1 970 1 968 1 949 1
   942 1 936 1 933 1 915 1 912 1 908 1 867 1 862 1 858 1 856 1 854 1 853
   1 851 2 848 2 844 2 837 1 830 1 829 1 819 1 798 4 797 1 796 1 791 1
   787 2 785 3 784 3 760 1 750 2 745 1 733 1 715 2 700 1 699 1 690 1 684
   2 667 1 654 2 645 1 636 3 632 2 630 1 628 1 621 1 614 2 581 2 576 1
   571 1 569 3 568 1 527 3 513 1 473 1 469 1 462 1 429 2 422 1 330 1 329
   1 322 1 321 1 317 1 314 1 282 1 277 2 266 1 209 1 207 1 203 1 199 2
   186 1 184 1 170 2 169 1 147 1 123 1 122 1 119 1 116 1 114 2 113 1 112
   1 111 3 109 4 108 4 106 2 105 1 96 1 94 1 91 1 89 1 84 1 83 4 82 1 81
   2 79 1 78 1 77 3 76 1 75 8 74 4 69 2 68 1 67 5 66 1 65 4 64 3 63 3 62
   1 60 1 59 3 58 2 54 1 46 1 39 3 29 1 28 5 27 1 25 1 23 1 12 1 10 1 7 1
   6 1
gepetto 8.33 1410 2
madeira 4.22 4137 1 4135 1 4044 1 4030 1 3927 1 3894 1 3868 1 3867 1 3863
   2 3818 1 3802 1 3756 1 3753 3 3711 1 3690 2 3689 1 3687 1 3654 3 3598
   4 3596 1 3587 7 3442 1 3329 1 3319 1 3307 1 3277 1 3276 2 3265 1 3199
   1 3191 1 3178 1 3173 1 3137 1 3089 2 2902 2 2881 1 2807 1 2780 1 2630
   1 2564 1 2406 1 2301 1 1885 1 1844 1 1814 3 1813 1 1812 5 1807 2 1802
   3 1630 1 1562 1 1559 1 1558 2 1429 2 1415 1 1410 1 949 1 492 1 465 1
   238 2 137 2
pinoquio 7.23 2943 1 1510 1 1410 3
teatro 3.54 4034 3 4031 1 4018 1 3954 2 3950 1 3926 1 3904 2 3903 1 3856
   1 3774 1 3749 1 3688 1 3616 1 3545 2 3544 1 3526 1 3521 1 3504 1 3496
   1 3495 1 3400 1 3388 1 3322 3 3305 1 3298 1 3288 6 3266 1 3254 7 3232
   4 3217 1 3215 1 3213 2 3212 1 3187 1 3184 2 3164 2 3162 1 3146 3 3145
   2 3117 1 3112 1 3108 1 3101 1 3093 1 3073 1 3064 1 3062 2 3046 1 3041
   1 3015 1 3010 1 2999 1 2907 2 2769 1 2682 2 2680 1 2652 9 2651 9 2647
   2 2621 1 2618 1 2561 5 2530 1 2485 5 2481 1 2477 4 2465 1 2440 1 2412
   1 2385 1 2328 1 2325 1 2322 1 2317 5 2312 1 2265 1 2252 1 2251 4 2250
   8 2153 1 2070 1 1786 1 1755 9 1740 1 1734 1 1727 6 1710 3 1690 5 1688
   1 1683 2 1681 1 1680 3 1676 2 1648 1 1614 1 1605 1 1601 2 1593 1 1590
   2 1587 3 1578 1 1577 1 1552 1 1538 1 1534 1 1507 1 1479 1 1474 1 1467
   1 1464 1 1414 4 1410 2 1409 1 1408 1 1407 2 1101 1 1031 1 872 1 842 1
   741 2 339 1
```

Estratégia de indexação BIGR:

```
(conteúdo: descritor IDF documento freqüência ...)
boneco=madeira 8.33 1410 1
criacao=gepetto 8.33 1410 1
madeira=pinoquio 8.33 1410 1
teatro 5.93 3904 1 3903 1 3288 1 3254 1 3015 1 2561 1 2328 1 2322 1 1578
   1 1577 1 1410 1
```

Estratégia de indexação SINN:

```
(conteúdo: descritor IDF documento freqüência ...)
boneco=de=madeira=pinoquio 8.33 1410 1
gepetto 8.33 1410 2
teatro 4.81 3954 1 3926 1 3904 2 3903 1 3856 1 3504 1 3495 1 3388 1 3288
   2 3254 4 3213 1 3212 1 3146 2 3064 1 3015 1 2769 1 2652 1 2561 1 2477
```

```
   1 2328 1 2322 1 2317 2 2252 1 2153 1 2070 1 1690 1 1683 1 1680 1 1578
   1 1577 1 1474 1 1414 1 1410 1 842 1
```

Estratégia de indexação NMRL:

Arquivo de índice com RLBs de classificação:

```
(conteúdo:  identificador
              argumento_1
                 argumento_2 documento freqüência)
=
   gepetto
      criador 1410 8.5
```

Arquivo de índice com RLBs de associação:

```
(conteúdo:  identificador
              argumento_1
                 argumento_2 documento freqüência)
criacao
   gepetto
      boneco 1410 14.0
criacao.por
   boneco
      gepetto 1410 14.0
```

Arquivo de índice com RLBs de restrição:

```
(conteúdo:  identificador
              argumento_1
                 argumento_2 documento freqüência)
de
   boneco
      madeira 1410 8.5
   criacao
      boneco 1410 9.5
em
   boneco
      teatro 1410 10.0
por
   criacao
      gepetto 1410 9.5
```

Observação (sobre as RLBs): Note que as duas primeiras RLBs do tipo restrição exemplificadas são "de(boneco,madeira)" e "de(criacao,boneco)". O identificador "de" não é repetido por ser o mesmo nas duas relações.

Arquivo de índice com termos:

```
(conteúdo: descritor documento freqüência ...)
boneco  3980 4.5 3979 0.5 3967 3.0 3958 10.0 3886 1.5 3761 1.0 2273 3.0
   1926 14.0 1439 0.5 1426 2.5 1421 1.5 1418 17.0 1410 7.0 1407 2.0 1185
   1.5 933 2.5 357 6.0 310 8.0 304 2.5
criacao  4152 1.5 4128 2.5 4105 0.5 4090 1.5 4045 1.5 4030 2.5 3997 1.5
   3966 0.5 3964 1.5 3959 0.5 3946 1.5 3936 0.5 3926 3.5 3904 0.5 3895
   1.5 3894 1.5 3888 2.5 3885 0.5 3874 2.5 3861 0.5 3851 1.5 3843 6.5
   3825 1.5 3815 1.5 3808 1.5 3772 2.5 3766 0.5 3764 3.5 3761 1.5 3757
   1.5 3745 1.5 3733 1.5 3719 1.5 3718 1.5 3678 1.5 3669 0.5 3596 1.5
   3576 3.0 3565 1.5 3562 2.5 3542 2.5 3540 5.0 3539 3.0 3529 1.5 3521
   1.5 3518 1.5 3516 1.5 3510 1.5 3497 3.0 3495 1.5 3490 2.5 3473 4.0
   3445 6.0 3442 2.5 3436 4.0 3419 2.5 3405 2.5 3403 1.5 3401 0.5 3384
   3.0 3379 1.5 3374 4.0 3361 0.5 3349 0.5 3347 5.5 3344 1.5 3342 1.5
   3329 2.5 3312 1.5 3305 1.5 3289 0.5 3288 2.5 3282 1.5 3279 3.0 3277
   2.5 3272 1.5 3266 1.0 3246 0.5 3239 1.5 3225 1.5 3221 3.0 3219 2.0
```

3216 2.5 3213 2.5 3197 1.5 3192 1.5 3191 1.5 3190 3.0 3183 1.5 3178
1.5 3156 1.5 3145 2.5 3141 1.5 3134 1.5 3132 3.0 3130 3.5 3122 0.5
3082 1.5 3081 4.0 3079 1.5 3078 2.0 3076 2.5 3060 3.0 3056 2.5 3040
0.5 3039 2.0 3034 1.5 3024 0.5 3016 2.5 3013 1.5 3010 1.5 2990 2.5
2988 1.5 2964 2.5 2954 2.5 2937 1.5 2899 2.5 2898 5.5 2891 2.5 2849
1.5 2844 2.5 2840 0.5 2827 1.5 2826 3.5 2823 0.5 2818 5.0 2816 3.0
2813 3.5 2810 6.5 2805 1.5 2795 1.5 2792 2.5 2789 0.5 2787 2.0 2785
0.5 2763 2.5 2761 1.5 2760 1.5 2743 3.0 2730 1.5 2727 2.5 2721 6.0
2710 1.5 2705 1.5 2702 1.5 2696 2.5 2693 1.5 2691 3.0 2684 3.5 2678
4.5 2676 2.5 2675 1.5 2663 1.5 2653 3.0 2647 3.0 2646 0.5 2643 1.5
2640 1.5 2635 2.5 2622 1.5 2619 1.5 2614 0.5 2603 1.5 2594 2.5 2588
7.0 2587 3.5 2586 2.5 2583 2.0 2581 4.5 2572 0.5 2564 9.0 2561 1.5
2556 4.5 2552 5.0 2551 1.5 2548 2.5 2547 0.5 2546 4.0 2545 2.5 2538
4.0 2532 0.5 2530 4.0 2529 9.0 2528 1.5 2527 1.5 2521 1.5 2519 1.5
2518 1.5 2517 0.5 2504 2.0 2496 1.5 2493 0.5 2492 1.5 2491 1.5 2485
4.5 2479 2.0 2477 2.5 2476 2.0 2474 1.5 2468 3.0 2467 2.5 2464 2.5
2454 0.5 2452 2.0 2450 1.5 2449 4.0 2443 3.0 2438 0.5 2436 0.5 2434
4.0 2433 1.5 2432 1.5 2431 1.5 2421 1.5 2415 1.5 2412 1.5 2409 3.5
2404 3.5 2401 1.5 2399 1.5 2397 2.5 2388 1.5 2385 0.5 2379 4.0 2367
0.5 2357 2.0 2351 4.0 2350 6.0 2349 0.5 2348 1.5 2345 1.5 2337 0.5
2331 2.5 2330 1.5 2327 1.5 2326 1.5 2323 5.0 2321 0.5 2317 2.5 2305
2.5 2302 0.5 2301 0.5 2299 0.5 2298 2.5 2284 2.5 2275 1.5 2268 2.5
2265 1.5 2262 2.5 2260 1.5 2253 1.5 2249 8.5 2247 1.5 2246 1.5 2233
1.5 2220 2.5 2217 2.5 2212 6.0 2211 2.5 2208 6.0 2196 4.0 2187 2.5
2185 0.5 2181 1.5 2180 0.5 2174 1.5 2173 0.5 2172 6.5 2171 5.0 2170
1.5 2161 4.5 2159 2.5 2158 2.5 2143 1.5 2134 0.5 2133 2.5 2127 1.5
2126 2.5 2125 3.5 2092 1.0 2090 1.5 2085 2.5 2083 5.5 2080 3.0 2079
5.5 2068 2.5 2065 8.5 2049 4.0 2044 0.5 2015 4.0 2011 5.5 2010 7.0
2009 1.5 2001 3.0 1995 1.5 1989 2.5 1986 2.5 1983 3.0 1981 0.5 1975
1.5 1957 2.5 1955 3.0 1950 3.5 1945 6.5 1944 3.0 1942 2.5 1937 1.5
1934 1.5 1928 4.0 1927 1.5 1926 1.5 1925 2.5 1924 1.5 1918 3.5 1913
1.5 1904 1.5 1903 0.5 1885 1.5 1881 1.5 1880 4.5 1878 4.0 1869 1.5
1842 2.5 1841 2.5 1839 1.5 1835 1.5 1831 2.5 1821 2.5 1819 1.5 1816
1.5 1814 2.5 1809 2.5 1795 2.5 1789 1.5 1777 2.0 1767 1.5 1765 2.5
1754 1.5 1747 1.5 1746 2.5 1744 1.5 1743 7.5 1738 1.5 1737 1.5 1734
1.5 1727 5.5 1725 1.5 1724 2.5 1719 2.5 1695 1.5 1693 0.5 1686 1.5
1685 0.5 1675 3.5 1671 0.5 1661 1.5 1653 1.5 1652 1.5 1646 1.5 1640
3.0 1619 1.5 1614 2.5 1606 5.5 1605 1.5 1602 2.5 1593 6.0 1590 1.5
1578 6.0 1577 6.0 1572 2.0 1565 1.5 1558 4.0 1557 1.5 1552 1.5 1550
1.5 1542 1.5 1536 1.5 1533 1.5 1530 4.0 1528 1.5 1517 1.5 1510 2.5
1508 3.5 1507 2.5 1481 1.5 1472 2.5 1464 2.5 1439 0.5 1437 0.5 1433
1.5 1418 1.5 1416 1.5 **1410 2.5** 1399 1.5 1387 1.5 1374 2.5 1353 1.5
1348 3.5 1324 0.5 1317 2.5 1257 1.5 1251 2.5 1249 6.5 1245 0.5 1241
0.5 1221 1.5 1219 1.5 1204 0.5 1166 1.5 1157 1.5 1155 3.5 1116 2.5
1089 1.5 1070 1.5 1051 1.5 974 2.5 970 1.5 968 1.5 942 1.5 936 0.5 933
1.5 928 3.5 915 1.5 914 3.5 912 1.5 908 0.5 889 0.5 867 3.5 862 2.5
858 3.0 854 1.5 853 2.5 851 2.0 848 3.0 844 4.0 839 2.5 837 2.5 830
0.5 829 1.5 823 2.5 819 2.5 798 16.5 797 1.5 796 1.5 795 1.5 792 1.5
791 2.5 787 1.0 785 5.5 784 6.5 767 1.5 760 1.5 750 4.0 745 0.5 733
1.5 715 3.0 700 1.5 699 0.5 690 1.5 687 0.5 684 2.0 677 1.5 667 1.5
654 5.0 645 6.0 636 7.5 632 0.5 630 1.5 628 0.5 621 1.5 614 5.0 608
0.5 603 1.5 581 3.0 578 4.5 577 4.0 576 3.5 571 1.5 569 18.0 568 4.0
548 0.5 527 4.5 513 2.5 473 0.5 469 1.5 462 4.0 446 4.5 429 2.0 422
2.5 413 1.5 330 2.5 329 2.5 324 3.5 322 1.5 321 2.5 317 1.5 314 2.5
292 1.5 291 2.0 282 0.5 277 3.0 266 1.5 247 0.5 209 2.5 207 1.5 203
2.5 199 3.0 186 2.5 184 1.5 170 5.0 169 0.5 147 4.0 123 1.5 119 4.0
116 1.5 114 1.5 109 7.5 108 5.0 94 0.5 84 6.5 83 3.5 82 1.5 81 1.5 77
7.0 75 2.0 69 1.5 67 3.0 66 0.5 65 1.5 63 3.0 54 2.5 46 1.5 39 1.5 29
1.5 12 1.5 8 0.5 7 2.5 6 6.0

criador 4152 0.5 4105 0.5 4090 0.5 4040 1.5 4030 1.5 4011 1.5 3966 0.5
3964 0.5 3959 0.5 3957 4.5 3946 0.5 3936 0.5 3926 2.5 3904 0.5 3894
0.5 3888 1.5 3874 1.5 3851 0.5 3843 3.5 3825 0.5 3815 0.5 3808 1.5
3772 1.5 3765 3.5 3764 2.5 3761 0.5 3757 0.5 3745 0.5 3733 0.5 3719
1.5 3718 0.5 3679 2.5 3678 0.5 3669 0.5 3596 0.5 3576 1.0 3562 1.5
3542 1.5 3529 0.5 3521 0.5 3518 4.0 3516 0.5 3510 0.5 3497 1.0 3495
0.5 3490 1.5 3473 2.0 3445 1.5 3442 1.5 3436 2.0 3419 1.5 3405 1.5
3403 0.5 3401 0.5 3384 2.0 3379 0.5 3374 9.0 3361 0.5 3347 2.5 3344
0.5 3342 0.5 3329 1.5 3312 0.5 3289 0.5 3288 1.5 3282 0.5 3279 1.0

3277 1.5 3272 0.5 3246 0.5 3239 0.5 3225 0.5 3221 1.5 3219 0.5 3216
1.5 3213 1.5 3197 0.5 3192 0.5 3191 0.5 3190 1.0 3183 0.5 3178 0.5
3156 0.5 3145 1.5 3141 0.5 3134 0.5 3132 2.0 3130 1.5 3122 0.5 3081
0.5 3079 0.5 3078 1.0 3076 1.5 3060 0.5 3056 1.5 3039 1.5 3024 0.5
3016 1.5 3013 0.5 3010 0.5 3004 1.5 2990 1.5 2964 1.5 2954 1.5 2937
0.5 2899 1.5 2891 1.5 2849 0.5 2844 1.5 2827 0.5 2826 1.5 2823 0.5
2818 2.0 2816 1.0 2810 3.5 2805 0.5 2795 0.5 2787 0.5 2785 0.5 2763
1.5 2761 0.5 2760 0.5 2743 1.0 2727 1.5 2721 1.5 2710 0.5 2705 0.5
2702 0.5 2696 1.5 2693 0.5 2691 1.0 2680 2.5 2678 1.5 2676 1.5 2663
0.5 2653 0.5 2647 1.5 2640 0.5 2635 1.5 2622 0.5 2588 0.5 2587 0.5
2586 1.5 2583 0.5 2581 2.5 2572 0.5 2564 6.0 2552 3.0 2551 0.5 2548
1.5 2546 2.0 2545 1.5 2538 2.0 2530 0.5 2529 3.5 2528 0.5 2527 0.5
2521 0.5 2518 0.5 2504 1.0 2496 0.5 2491 0.5 2476 1.0 2474 0.5 2468
1.0 2467 1.5 2452 0.5 2449 2.0 2443 2.0 2436 0.5 2434 2.0 2433 0.5
2432 0.5 2431 0.5 2412 0.5 2409 2.5 2404 1.5 2401 0.5 2399 0.5 2397
1.5 2388 0.5 2385 0.5 2379 2.0 2367 0.5 2357 1.0 2351 0.5 2350 2.5
2349 0.5 2348 0.5 2345 0.5 2331 1.5 2327 1.0 2326 0.5 2323 1.5 2321
0.5 2317 1.5 2305 1.5 2301 2.0 2299 0.5 2298 1.5 2275 0.5 2268 1.5
2262 1.5 2260 0.5 2249 3.0 2236 1.5 2233 0.5 2220 1.5 2217 1.5 2212
3.0 2211 1.5 2208 1.5 2196 1.0 2187 1.5 2185 0.5 2181 0.5 2173 0.5
2172 3.5 2171 0.5 2170 0.5 2161 1.0 2159 1.5 2158 1.5 2143 0.5 2133
1.5 2127 0.5 2126 1.5 2125 1.5 2113 1.5 2085 1.5 2083 2.0 2080 1.0
2068 1.5 2065 4.0 2049 1.5 2044 0.5 2015 2.0 2011 1.5 2010 2.0 2009
0.5 2001 1.0 1995 0.5 1989 1.5 1986 1.5 1983 1.0 1982 1.5 1981 0.5
1975 0.5 1963 4.5 1957 1.5 1955 1.0 1950 1.0 1945 5.0 1944 2.0 1942
1.5 1937 0.5 1934 0.5 1928 1.5 1927 0.5 1926 0.5 1925 1.5 1924 2.0
1913 0.5 1904 0.5 1903 0.5 1885 0.5 1880 0.5 1878 2.0 1869 0.5 1842
1.5 1841 1.5 1839 0.5 1835 1.5 1831 1.5 1821 1.5 1819 0.5 1816 0.5
1814 1.5 1809 1.5 1795 0.5 1789 0.5 1777 0.5 1767 0.5 1754 0.5 1747
0.5 1746 1.5 1744 0.5 1743 4.5 1737 0.5 1727 2.5 1725 0.5 1724 1.5
1719 1.5 1708 1.5 1695 0.5 1693 0.5 1686 0.5 1685 0.5 1653 0.5 1652
0.5 1646 0.5 1643 0.5 1640 1.0 1619 0.5 1614 1.5 1606 2.5 1605 0.5
1602 1.5 1593 3.0 1590 0.5 1584 2.5 1578 2.0 1577 2.0 1572 0.5 1558
3.5 1557 0.5 1552 0.5 1533 0.5 1530 2.0 1528 0.5 1517 0.5 1510 1.5
1508 1.5 1507 1.5 1481 0.5 1472 1.5 1464 1.5 1439 0.5 1437 0.5 1433
0.5 1418 0.5 1416 0.5 **1410 1.5** 1399 0.5 1387 0.5 1374 1.5 1353 0.5
1348 1.5 1317 0.5 1257 0.5 1251 1.5 1249 3.5 1245 0.5 1221 0.5 1219
0.5 1204 0.5 1166 0.5 1157 0.5 1116 1.5 1089 0.5 1070 0.5 1051 0.5 974
1.5 970 0.5 968 0.5 942 0.5 936 0.5 933 0.5 915 0.5 912 0.5 908 0.5
867 2.5 862 1.5 858 0.5 856 1.5 854 0.5 853 1.5 851 1.0 848 1.0 844
3.0 837 1.5 830 0.5 829 0.5 819 1.5 798 4.0 797 0.5 796 0.5 791 1.5
787 1.0 785 2.5 784 3.5 760 0.5 750 2.0 745 0.5 733 0.5 715 1.0 700
0.5 699 0.5 690 0.5 684 1.0 667 0.5 654 2.0 645 0.5 636 4.5 632 1.0
630 0.5 628 0.5 621 0.5 614 2.0 581 1.0 576 1.5 571 0.5 569 2.5 568
0.5 527 1.5 513 1.5 473 0.5 469 0.5 462 0.5 429 1.0 422 1.5 330 1.5
329 1.5 322 0.5 321 1.5 317 0.5 314 1.5 282 0.5 277 1.0 266 0.5 209
1.5 207 0.5 203 1.5 199 1.0 186 1.5 184 0.5 170 3.0 169 0.5 147 0.5
123 0.5 122 0.5 119 0.5 116 1.5 114 2.0 113 0.5 112 3.5 111 10.5 109
3.0 108 9.0 106 2.0 105 1.5 96 0.5 94 0.5 91 1.5 89 3.5 84 0.5 83 8.0
82 0.5 81 3.0 79 1.5 78 3.5 77 4.5 76 3.5 75 23.0 74 11.0 69 5.0 68
2.5 67 8.5 66 0.5 65 7.0 64 9.5 63 2.5 62 0.5 60 1.5 59 5.5 58 2.0 54
1.5 46 0.5 39 7.5 29 0.5 28 10.5 27 0.5 25 0.5 23 0.5 12 0.5 10 1.5 7
1.5 6 1.5

gepetto **1410 7.0**

madeira 4137 0.5 4135 1.5 4044 2.5 4030 0.5 3927 0.5 3894 3.5 3868 0.5
3867 7.5 3863 2.0 3818 0.5 3802 1.5 3756 0.5 3753 4.5 3711 1.5 3690
2.0 3689 1.5 3687 1.5 3654 4.5 3598 4.0 3596 0.5 3587 10.5 3442 0.5
3329 1.5 3319 0.5 3307 0.5 3277 0.5 3276 5.0 3265 2.5 3199 0.5 3191
1.5 3178 1.5 3173 1.5 3137 0.5 3089 6.0 2902 3.0 2881 1.5 2807 0.5
2780 0.5 2630 2.5 2564 1.5 2406 0.5 2301 0.5 1885 0.5 1844 1.5 1814
4.5 1813 3.5 1812 11.0 1807 4.0 1802 4.5 1630 1.5 1562 0.5 1559 0.5
1558 10.0 1429 4.0 1415 2.5 **1410 1.5** 949 1.5 492 0.5 465 1.5 238 3.0
137 2.0

pinoquio 2943 1.5 1510 0.5 **1410 2.5**

teatro 4034 4.5 4031 1.5 4018 1.5 3954 5.0 3950 0.5 3926 0.5 3904 4.0
3903 0.5 3856 0.5 3774 2.5 3749 3.5 3688 1.5 3616 1.5 3545 1.0 3544
1.5 3526 0.5 3521 1.5 3504 4.5 3496 0.5 3495 0.5 3400 2.5 3388 0.5
3322 6.5 3305 0.5 3298 2.5 3288 12.5 3266 1.5 3254 8.5 3232 12.0 3217

1.5 3215 1.5 3213 2.0 3212 2.5 3187 1.5 3184 4.0 3164 4.0 3162 0.5
3146 6.5 3145 3.0 3117 1.5 3112 1.5 3108 0.5 3101 1.5 3093 1.5 3073
1.5 3064 0.5 3062 3.0 3046 0.5 3041 0.5 3015 0.5 3010 6.5 2999 1.5
2907 4.0 2769 3.5 2682 6.0 2680 2.5 2652 21.5 2651 10.5 2647 3.0 2621
2.5 2618 2.5 2561 5.5 2530 5.5 2485 10.5 2481 3.5 2477 8.0 2465 5.5
2440 1.5 2412 0.5 2385 1.5 2328 0.5 2325 3.5 2322 0.5 2317 8.5 2312
1.5 2265 1.5 2252 5.5 2251 6.0 2250 11.0 2153 0.5 2070 0.5 1786 4.5
1755 7.5 1740 1.5 1734 0.5 1727 8.0 1710 5.5 1690 7.5 1688 0.5 1683
5.0 1681 1.5 1680 6.5 1676 4.0 1648 1.5 1614 1.5 1605 3.5 1601 5.0
1593 0.5 1590 3.0 1587 2.5 1578 0.5 1577 0.5 1552 0.5 1538 1.5 1534
1.5 1507 1.5 1479 1.5 1474 0.5 1467 1.5 1464 0.5 1414 5.0 **1410 3.0**
1409 0.5 1408 0.5 1407 3.0 1101 0.5 1031 0.5 872 3.5 842 0.5 741 3.0
339 1.5

ANEXO D DIFERENÇAS EVIDENTES

O conceito de evidência, utilizado no cálculo do peso dos descritores, pode ser entendido melhor através dos seguintes exemplos. Considere os dois documentos a seguir, sendo cada um constituído, para simplificar a exemplificação, por uma sentença:

Documento A: "A fiel governanta, que trabalhou na casa de campo, e o mordomo fugiram".

Documento B: "O fiel mordomo, que fugiu para o campo, trabalhou na casa da governanta".

Considere, também, para que nenhum outro fator influencie o cálculo do peso, que os dois documentos têm comprimentos iguais à média da coleção e que todos os termos tem fator *IDF* = 1. Com essas condições, na Tabela D.1 são apresentados os pesos dos termos lematizados para os dois documentos utilizando a Equação 2, baseada em freqüência de ocorrência. Na Tabela D.2 são apresentados os pesos dos termos nominalizados e na Tabela D.3, os pesos das RLBs para os dois documentos utilizando as Equações 7, 8 e 9, baseadas em evidência.

Tabela D.1: Pesos dos descritores com cálculo baseado em freqüência de ocorrência

descritores		doc A ou B	doc A	doc B
		freqüência	$W_{t,1}$	$W_{t,2}$
termos lematizados	campo	1	1	1
	casa	1	1	1
	fiel	1	1	1
	fugir	1	1	1
	governanta	1	1	1
	mordomo	1	1	1
	trabalhar	1	1	1

Considere que, na aplicação das Equações 2 e 7, são usados os parâmetros k_1 e b com valores 1,2 e 0,75, respectivamente, conforme o que é usualmente adotado.

Tabela D.2: Pesos dos termos com cálculo baseado em evidência

descritores		documento A		documento B	
		evidência	$W_{t,A}$	evidência	$W_{t,B}$
termos nominalizados	campo	0,5	0,65	2,5	1,49
	casa	2,5	1,49	2,5	1,49
	fidelidade	1,5	1,22	1,5	1,22
	fuga	4,5	1,74	3,5	1,64
	fugitivo	2,5	1,49	1,5	1,22
	governanta	8,5	1,93	0,5	0,65
	mordomo	3,5	1,64	9,5	1,95
	trabalhador	1,5	1,22	1,5	1,22
	trabalho	3,5	1,64	3,5	1,64

Na Tabela D.1 não é possível distinguir termos mais ou menos representativos. Naturalmente, a freqüência de ocorrência restrita a um documento que contém apenas uma sentença pouco pode contribuir neste sentido. Por outro lado, basta uma sentença

para que o cálculo baseado em evidência consiga apontar os descritores mais importantes, conforme pode ser observado na Tabela D.2, no caso dos termos, e na Tabela D.3, no caso das RLBs.

Tabela D.3: Pesos das RLBs baseados em evidência

descritores		documento A		documento B	
		evidência	$W_{r,A}$	evidência	$W_{r,B}$
RLBs classificações	=(governanta,fugitivo)	11,0	1,98		
	=(governanta,trabalhador)	10,0	1,96		
	=(mordomo,fugitivo)	6,0	1,83	11,0	1,98
	=(mordomo,trabalhador)			11,0	1,98
RLBs restrições	de(fidelidade,governanta)	10,0	1,96		
	de(fidelidade,mordomo)			11,0	1,98
	de(fuga,governanta)	13,0	2,01		
	de(fuga,mordomo)	8,0	1,91	13,0	2,01
	de(trabalho,governanta)	12,0	2,00		
	de(trabalho,mordomo)			12,0	2,00
	em(trabalho,casa)	6,0	1,83	6,0	1,83
	para(fuga,campo)			6,0	1,83
	por(fuga,governanta)	13,0	2,01		
	por(fuga,mordomo)	8,0	1,91	13,0	2,01
	por(trabalho,governanta)	12,0	2,00		
	por(trabalho,mordomo)			12,0	2,00
RLBs associações	fuga.para(mordomo,campo)			12,0	2,00
	trabalho.em(governanta,casa)	11,0	1,98		
	trabalho.em(mordomo,casa)			12,0	2,00

Um texto pode ser representado como uma estrutura da dados [GON 2001c, GON 2003]. De acordo com os pesos baseados em evidência, representações dos documentos A e B na forma de grafos são apresentadas, respectivamente, na Figura D.1 e na Figura D.2.

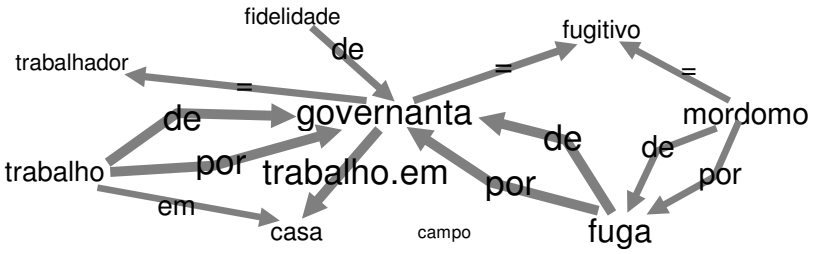

Figura D.1: Representação do documento A em grafo

Nesses grafos, os nodos são termos nominalizados e os arcos são RLBs. A espessura das setas e o tamanho dos caracteres são proporcionais aos pesos dos descritores para simular a representatividade dos mesmos.

O termo "campo", no documento A, e o termo "governanta", no documento B, não estão presentes em nenhuma RLB porque, de acordo com o modelo TR+, estão envolvidos em relações não evidentes. Essas relações necessitam informações semânticas para serem identificadas. Por exemplo, em "trabalhou na casa de janeiro a maio" a segunda preposição ("de") não associa o que vem depois dela com "casa", ao contrário de "trabalhou na casa de campo" e de "trabalhou na casa da governanta". As

regras utilizadas para identificar as RLBs não detectam tais diferenças e, assim, não capturam dependências desse tipo.

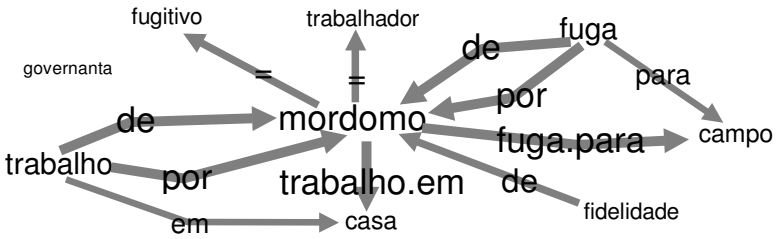

Figura D.2: Representação do documento B em grafo

Nos grafos apresentados ficam visíveis diferenças importantes entre os documentos A e B. Embora eles apresentem os mesmos termos que, por freqüência de ocorrência, não se destacam, a representatividade, com cálculo baseado em evidência, aponta diferenças. Por exemplo, a representatividade do termo "governanta" é grande no documento A e pequena no documento B. Desta forma, uma consulta com o termo "governanta" teria o documento A apontado como mais relevante.

As RLBs também têm representatividades que mostram diferenças entre os dois documentos. Uma consulta contendo "fuga de mordomo" recuperaria os dois documentos, tendo o documento B maior valor de relevância. Já "fuga de governanta" recuperaria o documento A como mais relevante.

ANEXO E TÓPICOS DE CONSULTA

São apresentados, neste Anexo, os 50 tópicos, para formulação de consultas, utilizados neste trabalho.

Tópico 1

Título: Abuso sexual

Descrição: Recuperar informação sobre abuso sexual sofrido por adulto ou criança.

Narrativa: Um documento relevante deve relatar ou comentar situação ou situações onde adultos ou crianças foram abusados sexualmente.

Tópico 2

Título: Acidente rodoviário

Descrição: Recuperar informação sobre acidente ocorrido em rodovia.

Narrativa: Um documento relevante deve relatar ou comentar acidente ocorrido em rodovia envolvendo qualquer tipo de dano.

Tópico 3

Título: Almoço

Descrição: Recuperar informação sobre almoço.

Narrativa: Um documento relevante deve relatar ou comentar encontros de pessoas para almoço ou informar sobre pratos servidos em um almoço ou, ainda, sobre preços ou locais deste tipo de refeição.

Tópico 4

Título: Animação

Descrição: Recuperar informação sobre animação de pessoas, desenhos ou bonecos.

Narrativa: Um documento relevante deve relatar ou comentar o ato de alguém se animar ou animar outra pessoa, ou descrever ou comentar a arte de animação de desenhos ou bonecos envolvendo computação gráfica ou qualquer tipo de técnica em produção cinematográfica, de televisão ou alguma mídia digital.

Tópico 5

Título: Bolsa de valores

Descrição: Recuperar informação sobre bolsa de valores.

Narrativa: Um documento relevante deve relatar ou comentar situações que envolvam instituição destinada a operar com ações de companhias ou outros títulos de crédito.

Tópico 6

Título: Campanha eleitoral de Lula

Descrição: Recuperar informação sobre a campanha para eleição presidencial de Luis Inácio Lula da Silva.

Narrativa: Um documento relevante deve relatar ou comentar situações sobre a campanha eleitoral de Luis Inácio Lula da Silva para presidente do Brasil.

Tópico 7

Título: Caso de cólera

Descrição: Recuperar informação sobre ações de combate ou efeitos de caso de cólera

Narrativa: Um documento relevante deve relatar ou comentar ações de combate ou efeitos de caso (ou casos) de doença infecciosa aguda, contagiosa, que pode manifestar-se sob forma epidêmica, conhecida pelo nome de “cólera”.

Tópico 8

Título: Certificação

Descrição: Recuperar informação sobre certificação.

Narrativa: Um documento relevante deve relatar ou comentar fatos que envolvam atribuição de algum tipo de certificado a alguém ou a algum produto ou a alguma empresa.

Tópico 9

Título: Cinema brasileiro

Descrição: Recuperar informação sobre o cinema brasileiro.

Narrativa: Um documento relevante deve relatar ou comentar fatos que envolvam o cinema brasileiro, ou seja, filmes produzidos no Brasil com artistas, diretores e recursos nacionais.

Tópico 10

Título: Cirurgia

Descrição: Recuperar informação sobre cirurgia médica.

Narrativa: Um documento relevante deve relatar ou comentar intervenção cirúrgica tanto com objetivo de diagnóstico quanto de cura de alguma doença.

Tópico 11

Título: Dança

Descrição: Recuperar informação sobre dança.

Narrativa: Um documento relevante deve relatar ou comentar fatos relacionados a qualquer tipo de dança, seja clássica, moderna, folclórica ou outro tipo, seja profissional ou realizada por divertimento.

Tópico 12

Título: Deputado federal

Descrição: Recuperar informação sobre ações ou características de algum deputado federal.

Narrativa: Um documento relevante deve relatar ou comentar situação envolvendo algum deputado federal ou descrever características de algum deputado federal.

Tópico 13

Título: Desemprego

Descrição: Recuperar informação sobre causas ou efeitos de desemprego.

Narrativa: Um documento relevante deve relatar ou comentar situações decorrentes de desemprego ou fatos que levam alguém a perder emprego.

Tópico 14

Título: Digitalização

Descrição: Recuperar informação sobre o processo de digitalização de documentos.

Narrativa: Um documento relevante deve descrever ou comentar dispositivos, técnicas ou efeitos de digitalização de textos, imagens ou qualquer outro tipo de documento.

Tópico 15

Título: Distribuição de renda

Descrição: Recuperar informação sobre distribuição de renda.

Narrativa: Um documento relevante deve relatar ou comentar efeitos ou benefícios da distribuição de renda, ou ações destinadas à sua promoção.

Tópico 16

Título: Drible

Descrição: Recuperar informação sobre situação em que tenha ocorrido drible.

Narrativa: Um documento relevante deve descrever ou comentar situação ou efeito de situação em que tenha ocorrido drible, em contexto esportivo ou não, como em “driblar a concorrência”.

Tópico 17

Título: Escola de samba

Descrição: Recuperar informação sobre ações ou características de uma escola de samba.

Narrativa: Um documento relevante deve descrever características de uma escola de samba, ou relatar ou comentar situação ou evento em que uma escola de samba tenha se envolvido.

Tópico 18

Título: Exportação

Descrição: Recuperar informação sobre exportação de algum produto.

Narrativa: Um documento relevante deve relatar ou comentar fato envolvido com exportação de algum produto.

Tópico 19

Título: Financiamento agrícola

Descrição: Recuperar informação sobre financiamento agrícola

Narrativa: Um documento relevante deve relatar ou comentar medidas destinadas a promover financiamento agrícola, ou efeitos deste tipo de financiamento.

Tópico 20

Título: Franquia

Descrição: Recuperar informação sobre franquia.

Narrativa: Um documento relevante deve relatar ou comentar fato envolvido com franquia de serviços ou produtos.

Tópico 21

Título: Globalização

Descrição: Recuperar informação sobre causas e efeitos de globalização.

Narrativa: Um documento relevante deve relatar ou comentar causas ou efeitos da globalização, como crescente integração de vários países, em termos de economias, culturas e outros aspectos.

Tópico 22

Título: Guerra do Golfo

Descrição: Recuperar informação sobre a Guerra do Golfo

Narrativa: Um documento relevante deve relatar aspectos ou comentar causas e conseqüências da Guerra do Golfo.

Tópico 23

Título: Hotel

Descrição: Recuperar informação sobre hotel.

Narrativa: Um documento relevante deve descrever um hotel ou vários hotéis, ou relatar ou comentar preços, promoções, instalações e outros aspectos característico de um ou vários hotéis.

Tópico 24

Título: Imóvel usado

Descrição: Recuperar informação sobre imóvel usado.

Narrativa: Um documento relevante deve descrever características de um ou mais imóveis usados, ou relatar ou comentar transação comercial ou reforma de imóvel usado.

Tópico 25

Título: Impressora

Descrição: Recuperar informação sobre impressora.

Narrativa: Um documento relevante deve descrever características do periférico de computador conhecido como impressora, ou relatar ou comentar fato envolvendo impressora como elemento principal.

Tópico 26

Título: Informatização

Descrição: Recuperar informação sobre informatização.

Narrativa: Um documento relevante deve relatar ou comentar causas, dificuldades, efeitos de informatização de empresa ou serviço, ou descrever alguma informatização realizada.

Tópico 27

Título: Instrumento musical

Descrição: Recuperar informação sobre instrumento musical.

Narrativa: Um documento relevante deve descrever características ou relatar o histórico ou a origem de um instrumento musical, ou explicar a contribuição de um instrumento musical em uma orquestra, banda ou outro tipo de grupo musical.

Tópico 28

Título: Kit multimídia

Descrição: Recuperar informação sobre características de um kit multimídia.

Narrativa: Um documento relevante deve descrever um kit multimídia ou comentar as vantagens de seu uso.

Tópico 29

Título: Leilão de gado

Descrição: Recuperar informação sobre a ocorrência e objetivo de um leilão de gado

Narrativa: Um documento relevante deve relatar ou comentar a ocorrência de um leilão de gado, ou descrever o tipo de animal leiloado ou transações realizadas.

Tópico 30

Título: Liderança de campeonato

Descrição: Recuperar informação sobre liderança de esportista ou equipe em campeonato.

Narrativa: Um documento relevante deve relatar ou comentar causas ou efeitos da liderança de esportista

ou equipe em campeonato que disputa.

Tópico 31

Título: Medalha de ouro

Descrição: Recuperar informação sobre disputa ou obtenção de medalha de ouro.

Narrativa: Um documento relevante deve relatar ou comentar a disputa por medalha de ouro ou a obtenção da mesma.

Tópico 32

Título: Merenda escolar

Descrição: Recuperar informação sobre distribuição de merenda escolar.

Narrativa: Um documento relevante deve relatar ou comentar causas de insucesso, ações para promover ou características da distribuição de merenda escolar, ou apontar responsáveis.

Tópico 33

Título: Mutuário

Descrição: Recuperar informação sobre mutuário.

Narrativa: Um documento relevante deve relatar ou comentar situações envolvendo mutuário, onde este é participante principal.

Tópico 34

Título: Nudismo

Descrição: Recuperar informação sobre local ou prática de nudismo.

Narrativa: Um documento relevante deve relatar ou comentar prática de nudismo, ou descrever características de local desta prática.

Tópico 35

Título: Passeio de barco

Descrição: Recuperar informação sobre passeio de barco.

Narrativa: Um documento relevante deve relatar ou comentar algum passeio onde o percurso tenha sido realizado através de barco.

Tópico 36

Título: Pastilha de freio

Descrição: Recuperar informação sobre pastilha de freio.

Narrativa: Um documento relevante deve descrever vantagens ou desvantagens de algum tipo ou marca de pastilha de freio, ou relatar ou comentar situação onde pastilha de freio tem participação importante.

Tópico 37

Título: Pintura restaurada

Descrição: Recuperar informação sobre pintura restaurada.

Narrativa: Um documento relevante deve relatar ou comentar restauração de pintura, ou descrever alguma pintura que foi, está sendo ou será restaurada.

Tópico 38

Título: Plano real

Descrição: Recuperar informação sobre o plano econômico denominado “Plano Real”.

Narrativa: Um documento relevante deve relatar ou comentar situações envolvendo o plano econômico conhecido como “Plano Real”, ou explicar causas e/ou conseqüências de sua implantação.

Tópico 39

Título: Pólo turístico

Descrição: Recuperar informação sobre pólo turístico.

Narrativa: Um documento relevante deve descrever um pólo turístico, ou relatar ou comentar situações características ou localizadas em algum pólo turístico.

Tópico 40

Título: Produtividade industrial

Descrição: Recuperar informação sobre produtividade industrial.

Narrativa: Um documento relevante deve relatar ou comentar ações que trazem aumento ou prejuízo para a produtividade industrial, ou relatar ou comentar efeitos do aumento ou da diminuição da produtividade industrial.

Tópico 41

Título: Projeto arquitetônico

Descrição: Recuperar informação sobre projeto arquitetônico.

Narrativa: Um documento relevante deve descrever aspectos de um projeto arquitetônico, ou comentar sobre os responsáveis, ou relatar ou comentar situações características de um projeto arquitetônico, tanto em relação à sua fase de realização, quanto aos seus efeitos depois de realizado.

Tópico 42

Título: Propaganda eleitoral gratuita

Descrição: Recuperar informação sobre propaganda eleitoral gratuita.

Narrativa: Um documento relevante deve relatar ou comentar situações envolvendo propaganda eleitoral realizada em horário eleitoral gratuito.

Tópico 43

Título: Publicação eletrônica

Descrição: Recuperar informação sobre publicação eletrônica.

Narrativa: Um documento relevante deve descrever o resultado ou características de uma publicação eletrônica ou características de dispositivo ou processo específico para publicação em meio eletrônico.

Tópico 44

Título: Reajuste salarial

Descrição: Recuperar informação sobre reajuste salarial.

Narrativa: Um documento relevante deve relatar ou comentar situações ou campanhas envolvendo tratativas para reajuste de salário, ou comentar reajustes salariais efetivados ou não obtidos.

Tópico 45

Título: Reciclagem de lixo

Descrição: Recuperar informação sobre reciclagem de lixo.

Narrativa: Um documento relevante deve descrever processos para reciclagem de lixo, ou relatar ou comentar medidas para promover a reciclagem de lixo, ou informar sobre responsáveis ou sobre locais onde ocorre ou ocorrerá.

Tópico 46

Título: Seleção brasileira de futebol

Descrição: Recuperar informação sobre a seleção brasileira de futebol.

Narrativa: Um documento relevante deve relatar ou comentar situações, disputas ou participantes da seleção brasileira de futebol.

Tópico 47

Título: Treino oficial

Descrição: Recuperar informação sobre treino oficial.

Narrativa: Um documento relevante deve relatar ou comentar situações, participantes ou resultados de um treino oficial de competição automobilística.

Tópico 48

Título: Uno Mille

Descrição: Recuperar informação sobre Uno Mille.

Narrativa: Um documento relevante deve descrever versões do automóvel Uno Mille, ou relatar ou comentar situações onde um veículo dessa marca tem participação importante.

Tópico 49

Título: Vestibular

Descrição: Recuperar informação sobre concurso vestibular.

Narrativa: Um documento relevante deve relatar ou comentar situação característica ou peculiar de um concurso vestibular, de seus participantes ou de seus organizadores.

Tópico 50

Título: Viagem de carro

Descrição: Recuperar informação sobre viagem de carro.

Narrativa: Um documento relevante deve relatar ou comentar situação envolvendo uma viagem realizada através de carro, ou a preparação do mesmo para viagem.

ANEXO F DOCUMENTOS JULGADOS RELEVANTES

São apresentadas, neste Anexo, as listagens dos documentos julgados relevantes para cada tópico de consulta que consta do Anexo E.

Tópico 1: Abuso sexual
Documentos: 268, 271, 274, 301, 302, 303, 313, 357, 358, 396, 401, 407, 408, 436, 437, 449, 477, 478, 479, 510, 538, 539, 2832
Tópico 2: Acidente rodoviário
Documentos: 139, 260, 307, 308, 410, 452
Tópico 3: Almoço
Documentos: 321, 869, 898, 1006, 1273, 2532, 2872, 3014, 3029, 3044, 3090, 3157, 3181, 3199, 3285, 3313, 3326, 3442, 3648, 3659, 3763, 3786, 3795, 3875, 3879
Tópico 4: Animação
Documentos: 596, 697, 876, 931, 1182, 1226, 1312, 1384, 1401, 1402, 1418, 1498, 1581, 1643, 1646, 1690, 1714, 1910, 1925, 1950, 1956, 1976, 1996, 2006, 2022, 2145, 2152, 2157, 2161, 2169, 2170, 2171, 2172, 2173, 2187, 2199, 2208, 2211, 2212, 2234, 3008, 3021, 3064, 3141, 3472, 3521, 3627, 3737, 3760
Tópico 5: Bolsa de valores
Documentos: 599, 626, 651, 2696, 2790, 2915, 2928, 3771
Tópico 6: Campanha eleitoral de Lula
Documentos: 872, 882, 912, 926, 927, 929, 933, 934, 957, 981, 2275, 2911, 2927, 2931, 2972, 2982, 3244, 3444, 3505
Tópico 7: Caso de cólera
Documentos: 262, 305, 354, 400, 444, 445, 544, 580, 581, 1317
Tópico 8: Certificação
Documentos: 701, 3571
Tópico 9: Cinema brasileiro
Documentos: 1675, 1704, 1717, 1737, 3026, 3035, 3141, 3271, 3956
Tópico 10: Cirurgia
Documentos: 797, 798, 809, 810, 1324, 2842, 2864, 3480
Tópico 11: Dança
Documentos: 275, 464, 1414, 1432, 1436, 1568, 1571, 1572, 1587, 1605, 1606, 1614, 1727, 2616, 2796, 3030, 3205, 3225, 3254, 3272, 3437, 3471, 3684, 3835, 3857, 3858, 3904
Tópico 12: Deputado federal
Documentos: 117, 255, 256, 257, 285, 292, 293, 296, 326, 346, 347, 348, 350, 470, 504, 872, 896, 902, 916, 919, 930, 931, 1662, 2331, 2334, 2335, 2367, 2579, 2581, 3341, 3378, 3429, 3430, 3431, 3432, 3443, 3972
Tópico 13: Desemprego
Documentos: 165, 603, 855, 946, 1713, 2350, 2409, 2690, 2693, 2710, 2903, 2965, 3079, 3852, 3890
Tópico 14: Digitalização
Documentos: 1981, 2023, 2047, 2065, 2079, 2091, 2122, 2125, 2195, 2213, 2225, 2540
Tópico 15: Distribuição de renda
Documentos: 1557, 2236, 2320, 2654, 2961
Tópico 16: Drible
Documentos: 455, 1162, 1197, 1281, 1366, 1374, 1375, 3064, 3474
Tópico 17: Escola de samba
Documentos: 93, 468, 470, 2676, 3295, 3300, 3303, 3317, 3319, 3778, 3959, 4011, 4023
Tópico 18: Exportação
Documentos: 14, 33, 37, 46, 47, 53, 57, 73, 84, 88, 97, 98, 100, 104, 165, 207, 224, 335, 572, 574, 584, 588, 589, 590, 599, 601, 603, 608, 613, 618, 621, 622, 626, 630, 635, 651, 779, 823, 830, 1278, 1635, 1843, 2064, 2073, 2172, 2450, 2513, 2582, 2588, 2591, 2594, 2643, 2725, 2757, 2792, 2814, 2826, 2877, 3175, 3176, 3541, 3637, 3685, 3687, 3756, 3850, 3851, 3975, 4143, 4150
Tópico 19: Financiamento agrícola
Documentos: 1, 6, 14, 29, 30, 87, 88, 91, 97, 101, 104
Tópico 20: Franquia
Documentos: 535, 679, 688, 688, 712, 733, 736, 751, 795, 1125, 1770, 1772, 1773, 1775, 1776, 3518, 3522, 3534, 3548, 3560, 3561, 3562, 3563, 3825
Tópico 21: Globalização
Documentos: 147, 727, 2409, 2584, 2587, 2687
Tópico 22: Guerra do Golfo
Documentos: 130, 132, 844, 2195, 2241, 2855, 2891, 2893, 3160, 3279
Tópico 23: Hotel
Documentos: 466, 595, 614, 894, 1134, 1166, 1167, 1574, 2542, 2740, 2742, 3104, 3262, 3390, 3592, 3594, 3652, 3658, 3663, 3682, 3685, 3687, 3691, 3693, 3694, 3695, 3696, 3701, 3708, 3710, 3713, 3725, 3733, 3752, 3758, 3763, 3764, 3774, 3778, 3779, 3782, 3787, 3793, 3795, 3796, 3814, 3816, 3832, 3842, 3870, 3874, 3875, 3877, 3878, 3879, 3899, 3906, 3909, 3916, 3922
Tópico 24: Imóvel usado
Documentos: 1797, 1798, 1799, 1800, 1801, 1810

Tópico 25: Impressora
Documentos: 1500, 1958, 1964, 1985, 1986, 1993, 2018, 2020, 2021, 2033, 2037, 2039, 2052, 2060, 2061, 2062, 2067, 2083, 2084, 2106, 2107, 2109, 2111, 2120, 2131, 2151, 2160, 2169, 2183, 2205, 2213, 2217, 2219, 3539, 3555
Tópico 26: Informatização
Documentos: 567, 785, 1765, 1794, 1795, 1933, 1936, 2005, 2078, 2650, 3054
Tópico 27: Instrumento musical
Documentos: 1432, 1507, 1515, 1529, 1539, 1586, 1600, 1640, 1727, 2129, 2130, 2208, 2222, 2226, 3131, 3830, 3855
Tópico 28: Kit multimídia
Documentos: 1505, 2050, 2139, 2140, 2209, 2214, 2218, 2219, 2221
Tópico 29: Leilão de gado
Documentos: 8, 24, 28, 39, 60, 62, 63, 65, 67, 79, 106, 111, 112
Tópico 30: Liderança de campeonato
Documentos: 1009, 1010, 1036, 1059, 1060, 1071, 1098, 1099, 1109, 1117, 1170, 1179, 1194, 1195, 1200, 1228, 1241, 1252, 1263, 1266, 1277, 1292, 1293, 1301, 1340, 1350, 1367
Tópico 31: Medalha de ouro
Documentos: 864, 1382, 3371, 3420
Tópico 32: Merenda escolar
Documentos: 273, 542, 1642
Tópico 33: Mutuário
Documentos: 646, 1805, 1822
Tópico 34: Nudismo
Documentos: 3624, 3666
Tópico 35: Passeio de barco
Documentos: 3442, 3625, 3645, 3653, 3655, 3698, 3705, 3711, 3735, 3782, 3832, 3841, 3910
Tópico 36: Pastilha de freio
Documentos: 818, 3293, 4099, 4110
Tópico 37: Pintura restaurada
Documentos: 2264, 2778, 2787
Tópico 38: Plano real
Documentos: 71, 72, 76, 100, 101, 103, 120, 136, 179, 187, 191, 203, 204, 208, 209, 210, 216, 234, 235, 589, 591, 594, 613, 617, 624, 627, 628, 660, 872, 882, 903, 926, 945, 946, 965, 967, 1816, 1831, 1866, 1867, 2379, 2380, 2381, 2409, 2410, 2430, 2557, 2990, 3170, 3175, 3176, 3187, 3194, 3739
Tópico 39: Pólo turístico
Documentos: 3661, 3732, 3757, 3822
Tópico 40: Produtividade industrial
Documentos: 29, 104, 593, 701, 788, 820, 830, 2690, 3567
Tópico 41: Projeto arquitetônico
Documentos: 1558, 1771, 1783, 1832, 1900, 2647, 3757
Tópico 42: Propaganda eleitoral gratuita
Documentos: 876, 885, 889, 890, 892, 915, 928, 931, 932, 933, 962, 963, 979, 981
Tópico 43: Publicação eletrônica
Documentos: 1498, 1505, 1523, 2070, 2071, 2074, 2222, 2224
Tópico 44: Reajuste salarial
Documentos: 129, 130, 133, 134, 168, 174, 177, 178, 191, 195, 214, 216, 218, 219, 222, 226, 248, 250, 600, 646, 734, 752, 770, 946, 2954
Tópico 45: Reciclagem de lixo
Documentos: 289, 555, 571, 572, 1438
Tópico 46: Seleção brasileira de futebol
Documentos: 994, 1005, 1006, 1011, 1024, 1025, 1061, 1131, 1134, 1137, 1138, 1153, 1160, 1161, 1164, 1166, 1167, 1172, 1186, 1189, 1215, 1216, 1242, 1253, 1256, 1257, 1267, 1273, 1286, 1287, 1288, 1289, 1292, 1293, 1337, 1339, 1340, 1342, 1343, 1344, 1347, 1354, 1366, 1367, 1379, 1388, 1392, 1393, 1394, 1395, 1424, 1448, 1452, 1476, 2024, 2230, 2232
Tópico 47: Treino oficial
Documentos: 1036, 1149, 1176, 1177, 1200, 1203, 1206, 1207, 1252, 1311, 3239
Tópico 48: Uno Mille
Documentos: 4046, 4047, 4049, 4054, 4057, 4108, 4127
Tópico 49: Vestibular
Documentos: 154, 512, 1217, 1301, 1313, 1484, 1485, 1487, 1488, 1489, 1493, 1495, 1497, 2512, 3445, 3489, 3671, 3672, 3673, 3674, 3675, 3676, 4031
Tópico 50: Viagem de carro
Documentos: 1332, 1364, 1372, 3783, 3788, 3790, 4046, 4088, 4099, 4100