UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Text Mining, um estudo a partir da rede social *Twitter*

Autor: Gabriel Luiz Andriotti da Silva

Orientadora: Professora Dra. Lisiane Priscila Roldão Selau

Porto Alegre, 12 de julho de 2013.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Text Mining, um estudo a partir da rede social *Twitter*

Autor: Gabriel Luiz Andriotti da Silva

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professora Dra. Lisiane Priscila Roldão Selau
Professor Dr. Leandro Krug Wives

Porto Alegre, 12 de julho de 2013.

Dedico este trabalho aos meus pais, Sílvio e Jussânia.

Agradecimentos

A Deus, quem me criou e me deu vida e sabedoria pra viver de maneira honrosa e digna. Todas as coisas são Dele, por Ele e para Ele.

Aos meus pais, Sílvio e Jussânia, pelo amor, incentivo e paciência. Muito obrigado por serem referência pra minha vida em todo o tempo. Quando crescer, quero ser como vocês, pessoas íntegras, trabalhadoras, incansáveis, honestas e acima de tudo, pais de verdade que me ensinaram o caminho em que devo andar, do qual jamais vou me desviar. Amo muito vocês, obrigado por tudo.

Ao meu avô Derci, que sempre torceu muito por mim, e ficou muito emocionado quando eu passei no vestibular anos atrás e que infelizmente não poderá ver minha formatura pois faleceu. Quero honrá-lo aqui, pois mesmo sendo um homem simples, que nunca fez uma faculdade, incentivou muito meu pai a batalhar e estudar, legado que influenciou também minha vida.

Aos colegas da Estatística, Vanessa Schierholt e Eduardo Seligman, pela amizade e parceria. Nunca vou esquecer das inúmeras caronas e discussões mirabolantes voltando da faculdade. Obrigado por terem feito parte desse período tão importante da minha vida.

Aos professores do Departamento de Estatística por tudo que aprendi nesses anos juntos. Em especial o professor Beal, que nos ensinou probabilidade de um jeito memorável com seus exemplos nada convencionais. Obrigado por ter me dado conceito A, pois rendeu muitas histórias nos corredores da universidade, por ser praticamente impossível obter conceito máximo em tuas disciplinas. Também quero agradecer à professora Suzi, por ter paciência comigo sempre que eu chegava muito cansado nas aulas.

À minha orientadora, Lisiane, pela orientação neste trabalho e por ter me apoiado na escolha do assunto, mesmo sendo um tópico não visto durante a graduação.

Resumo

A cada dia, mais dados são armazenados por computadores cada vez mais poderosos e capazes de processar muita informação. Com isso, cresce a necessidade de utilizar toda essa informação no processo de tomada de decisão. Estima-se que 80% da informação que uma organização dispõe esteja no formato de texto. Além disso, o advento das redes sociais na *web* possibilitou um ambiente virtual em que milhões de usuários publicam dados pessoais, opiniões e preferências, tornando-se uma fonte preciosa de informação. Diante desse cenário, a descoberta de conhecimento em texto torna-se uma ferramenta importante para extrair informação a partir de bases textuais sem a necessidade de leitura. Nesse sentido, o objetivo deste trabalho é apresentar a abordagem de *Text Mining* com as fases que compõem o processo, desde a extração até a análise dos dados, e mostrar uma aplicação da técnica com dados do *Twitter*. Utilizando o *software* R, foram extraídas as publicações do *Twitter* da Zero Hora – jornal local – num período de quase 3 meses, e foi construída uma nuvem de palavras com os termos mais frequentes. E, através da técnica de *Clustering*, os *tweets* foram agrupados conforme seus assuntos.

Palavras-chave: *Text Mining*, *software* R, *clustering*, nuvem de palavras, *Twitter*.

Sumário

1. INTRODUÇÃO

Todos os dias é processada e armazenada uma grande quantidade de dados em todo o tipo de interação entre seres humanos e computadores. O avanço da tecnologia facilita cada vez mais o registro desse fluxo. Desde uma compra realizada em um estabelecimento comercial até uma postagem em redes sociais, um e-mail enviado para um amigo ou até mesmo a realização de um cadastro para assinatura de um produto, são exemplos de procedimentos comuns do dia a dia em que dados estão envolvidos.

Diante desse cenário, existe a necessidade de analisar esses dados, de explorá-los e extrair informação. As empresas, o governo e as pessoas de uma forma geral estão atentando para isso, a sociedade está ávida por informação. O processo decisório das instituições baseia-se cada vez mais em análise da informação proveniente da exploração de suas bases de dados.

Existem muitas técnicas e soluções para analisar dados estruturados, como o *Data Mining* – mineração de dados para descoberta de conhecimento a partir de dados numéricos. Porém, segundo Tan (1999), estima-se que aproximadamente 80% da informação que uma organização dispõe, está armazenada de forma não estruturada, em formato de texto, o que dificulta trabalhar com essa informação.

Nesse contexto, surgiu uma área derivada do *Data Mining*, chamada *Text Mining*, que extrai de textos, a partir de técnicas e processos, informação útil sem precisar de leitura prévia. Através do *Text Mining* é possível extrair informação desconhecida de grandes coleções de textos sem que haja necessidade da leitura humana. Podem ser descobertos padrões e relações entre os textos que seria muito difícil, ou quase impossível, encontrar lendo manualmente (TAN, 1999).

Entre outras fontes de textos, a internet como um todo, mas principalmente as redes sociais, fornece uma quantidade substancial de dados a respeito de comportamento, desejo e atitudes das pessoas.

O crescimento da internet impulsionou o *Text Mining*, sendo possível extrair informação valiosa dela para companhias e até mesmo para o governo. É possível extrair das páginas de redes sociais os dados das pessoas, bem como suas postagens e suas ações no ambiente virtual, e gerar informação a partir disso. Assim, pode-se, por exemplo, conhecer melhor as pessoas e oferecer a elas produtos e serviços específicos de acordo com o seu gosto. Uma empresa que oferece para seus clientes aquilo que eles gostam e querem comprar, com certeza terá muito mais sucesso em suas vendas. Pode-se também realizar controle de e-mails trocados por funcionários em uma empresa, conversações em *chats*

institucionais e manifestações de clientes em redes sociais reclamando ou elogiando uma marca, por exemplo.

Nesse sentido, o objetivo desse estudo é apresentar a abordagem de *Text Mining* com as fases que compõem o processo, desde a extração até a análise dos dados. Além desse objetivo geral, como objetivo específico, é feita uma aplicação de técnicas e algoritmos, apresentando um exemplo prático, a partir da rede social *Twitter*.

Este trabalho está estruturado em cinco seções. Nesta primeira seção de introdução o assunto do *Text Mining* é contextualizado, ressaltando-se sua importância e atualidade em função da grande massa de dados não numéricos disponíveis, o que serviu de motivação para este estudo. Na segunda parte será abordado de forma geral o *Text Mining* com suas técnicas, processos e alguns exemplos de aplicação. A terceira seção trata sobre os dados e *software* utilizados na aplicação que é apresentada na quarta seção, bem como as técnicas de *Text Mining* escolhidas. Na quarta seção é apresentada uma aplicação de *Text Mining*, com uso de textos de um jornal local no *Twitter*, mostrando cada uma de suas fases. Por fim, a última seção traz as principais conclusões deste estudo e sugestões de trabalhos futuros.

2. KDT - DESCOBERTA DE CONHECIMENTO EM TEXTO

KDT - *Knowledge Discovery in Text* ou Descoberta de Conhecimento em Texto e *Text Mining* são os nomes usados para descrever o conjunto de técnicas que extrai informação relevante a partir de bases textuais não-estruturadas (DIXON, 1997). O KDT combina técnicas e conhecimentos de diversos segmentos como Informática, Estatística, Linguística, Matemática e outras, tendo a capacidade de extrair conhecimento a partir de grandes coleções de textos.

A descoberta de conhecimento em texto deriva do KDD - Knowledge Discovery in Databases, também conhecido como *Data Mining*, porém no KDD a fonte usada para a exploração é estruturada. No KDT são aplicadas técnicas e processos a fim de estruturar os dados oriundos de textos para permitir a posterior mineração da informação.

Dixon (1997) sugere as seguintes etapas no processo de KDT:

1. Recuperação de Informação: Encontrar os textos que tenham informação relevante de acordo com o assunto pesquisado. Selecionar os documentos que contém a informação procurada.

2. Extração de Informação: Extrair informação dos documentos selecionados. Nesta fase o usuário indica itens relevantes para a construção do *template* – dados estruturados que podem ser usados em métodos de KDD tradicionais (WIVES, 2002).
3. Mineração: Uma vez que os textos selecionados foram convertidos em dados estruturados, pode-se aplicar técnicas do *Data Mining*. Nesta fase procura-se padrões e relacionamentos nos dados.
4. Interpretação: Interpretar os padrões encontrados na fase anterior. Análise das saídas, no formato de língua natural, do processo de mineração.

O passo 2 é muito importante, pois ele consiste em estruturar os textos para possibilitar a mineração no passo seguinte. Segundo Feldman (1995), antes de realizar qualquer tipo de descoberta de conhecimento em bases textuais, é necessário que a informação esteja estruturada.

Feldman (1995) afirma que é necessário decidir cuidadosamente como a estrutura será construída com os dados, levando em consideração que ela precisa ser simples para possibilitar a automatização, porém conter a informação necessária para responder as questões procuradas. Esse processo deve resultar em uma estrutura rica o bastante para se ter relevância na informação, mas sem custar muito para fazê-lo.

Depois de estruturadas, essas informações podem ser usadas por métodos tradicionais de descoberta de conhecimento. Nessa etapa é definido como e quais informações (palavras) serão extraídas. A informação importante é encontrada através de *tags* (marcas) sintáticas ou semânticas (WIVES, 2002).

Considerando, por exemplo, que datas têm o formato DD/MM/YY, podem ser criadas regras de extração para obter do texto as informações desejadas, como o ano em que algum fato ocorreu. Abaixo, um exemplo de um texto que passa pelo processo de extração de informação e é estruturado conforme a Figura 1.

Em março de 2012 ocorreu um protesto em Brasília, o qual envolveu 20.000 pessoas que lutavam contra a legalização do aborto.

Ano	2012
Lugar	Brasília
Instrumento	Protesto
Qtde Pessoas	20.000
Assunto	Aborto

Figura 1 - *Slot* preenchido de acordo com o texto acima

2.1 Técnicas de *Text Mining*

Os tipos de técnicas de *Text Mining* mais encontrados na literatura são: sumarização, classificação/categorização e *clustering* (WIVES, 2002).

2.1.1 Sumarização

Segundo Dixon (1997), a Sumarização consiste em extrair do texto um número pequeno de frases que possam resumir o assunto do documento original. São selecionadas as palavras e frases mais importantes do texto, ou conjunto de textos, a fim de não ser necessária a leitura prévia, e mesmo assim, obter a essência da mensagem.

Para facilitar a definição das regras de extração para sumarizar os documentos, é importante que os textos sejam semelhantes entre si. Por exemplo, se um conjunto de textos trata sobre um assunto específico, é mais fácil estipular os critérios de extração.

2.1.2 Classificação/Categorização

A Classificação serve para determinar a que classe um documento pertence, de acordo com as características definidas para cada classe. Enquanto a Categorização indica quais assuntos e temas estão contidos em um documento.

Os processos de Classificação/Categorização também podem ser usados para outros propósitos, como a extração de informações, leitores de *e-mail* ou notícias eletrônicas (WIVES, 2002). Segundo o autor, as técnicas de Classificação mais usadas são Regras de Inferência, Redes Neurais Artificiais, Método de Similaridade de Vetores ou Centróides, Árvores de Decisão e Classificadores de Bayes.

2.1.3 *Clustering*

Também conhecida como Agrupamento, esta técnica basicamente agrupa textos em classes de acordo com as características de cada documento, sem necessidade de alguma definição pelo usuário.

Segundo Wives (2002), o *Clustering* identifica co-relacionamentos e associações entre objetos para facilitar a identificação das classes. Assim, esta

técnica é muito útil para criar conjuntos de texto pelo assunto, sem precisar ter conhecimento prévio dos textos.

2.2 Aplicações de *Text Mining*

Nesta subseção são apresentados alguns exemplos de aplicação de *Text Mining* para exemplificar o quão útil a técnica pode ser, nas diversas áreas do conhecimento, para extrair informações valiosas a partir de bases textuais.

2.2.1 *Twitter* e Alegria

Cientistas da Universidade de Vermont descobriram, através de técnicas de *Text Mining*, uma forma de mensurar a alegria dos usuários da rede social *Twitter*. Os resultados do estudo foram publicados no artigo intitulado "*Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter* ".

Durante 3 anos, foram coletadas mais de 46 bilhões de palavras, escritas por 63 milhões de usuários da rede social espalhados pelo mundo. Para medir a alegria das pessoas em relação a cada palavra, os pesquisadores pagaram voluntários para votarem numa escala de 1 a 9, sendo 1 menos alegre e 9 mais, para as 10 mil palavras mais comuns da língua inglesa. Os cientistas de Vermont então usaram essa escala para classificar a alegria das postagens feitas no *Twitter* ao longo do período da pesquisa. Foi constatado que o nível de alegria está diminuindo, como mostra a Figura 2.

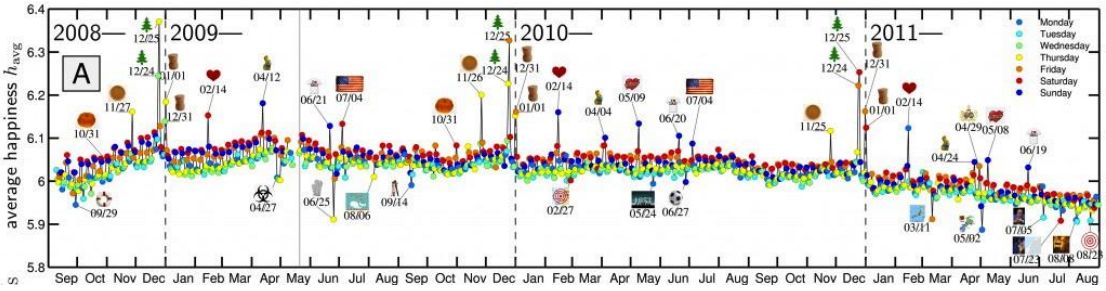

Figura 2 - Gráfico da média de alegria calculada no *Twitter*

Esse tipo de análise também pode ser usada com fontes de dados como jornais e sites de notícias, por exemplo, podendo assim mensurar de alguma forma o que se passa em uma certa localidade ao longo dos tempos. Pode ainda servir como complemento de pesquisas sociológicas.

2.2.2 Jornalismo

Em sua dissertação de mestrado, "*Descoberta de Conhecimento com o uso de Text Mining: Cruzando o Abismo de Moore"*, Edilberto Magalhães aplicou técnicas de *Text Mining* em uma empresa de Comunicação. Foram avaliadas as matérias publicadas e algumas estatísticas foram extraídas (SILVA, 2002).

Foi possível mensurar a produção mensal, quantidade média por dia de produção e tamanho médio de cada matéria. Contou-se a quantidade de palavras por texto, o que possibilitou calcular a média de palavras em cada texto por mês. Como no jornalismo, o tamanho de uma publicação é fator importante, e existem padrões para isso, esse foi um dos pontos relevantes do trabalho.

Com a contagem de palavras, através dos termos mais frequentes, permitiu-se descobrir quais os assuntos abordados pela empresa, como mostra a Figura 3. Também foi extraída dos textos a distribuição geográfica das matérias, possibilitando uma análise do alcance da cobertura jornalística realizada.

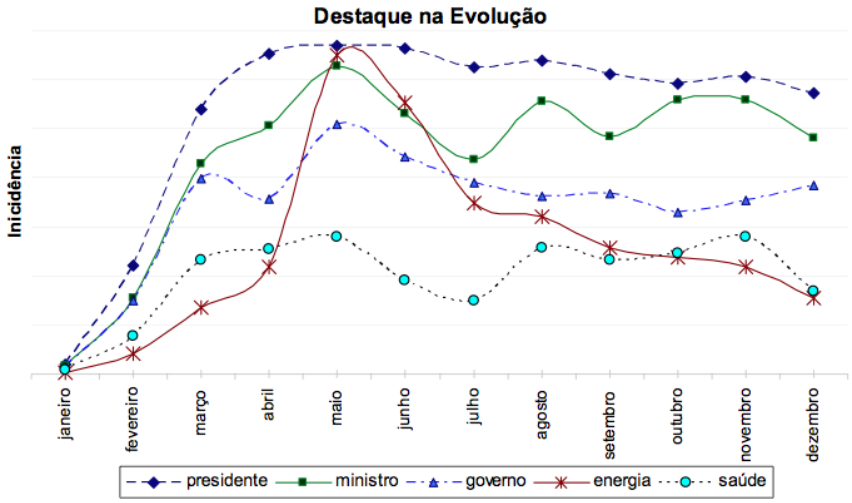

Figura 3 – Termos mais frequentes por mês

Silva (2002) obteve, como principal resultado do trabalho, a confirmação do papel social da empresa estudada junto à sociedade. Também foi possível aos administradores, através dos resultados, se certificarem do cumprimento de suas metas organizacionais, mostrando que o *Text Mining* pode ser muito útil na descoberta de conhecimento no meio empresarial.

2.2.3 Segurança Nacional

Alessandro Zanasi, professor na Universidade de Bologna, propõe, no seu artigo intitulado "*Virtual Weapons for Real Wars: Text Mining for National Security*"**,** o uso das técnicas de descoberta de conhecimento em texto para monitorar atividades dos terroristas e descobrir seus nomes e grupos de envolvimento. Usando, como base de dados, blogs, sites suspeitos, e-mails, mensagens trocadas em chats, entre outras fontes, é possível, através de ferramentas de extração de informações, controlar atividades possivelmente terroristas (ZANASI, 2008).

Através dessas ferramentas, pode-se acompanhar o uso de termos específicos, como por exemplo "*Al Qaeda*" ,"Hizbollah" e outros, como mostra a Figura 4, e com isso criar coleções de documentos que contenham estes termos, a fim de monitorar o que os terroristas estão publicando, ou o que está sendo publicado a respeito deles nos mais diversos meios, como os já citados no parágrafo anterior.

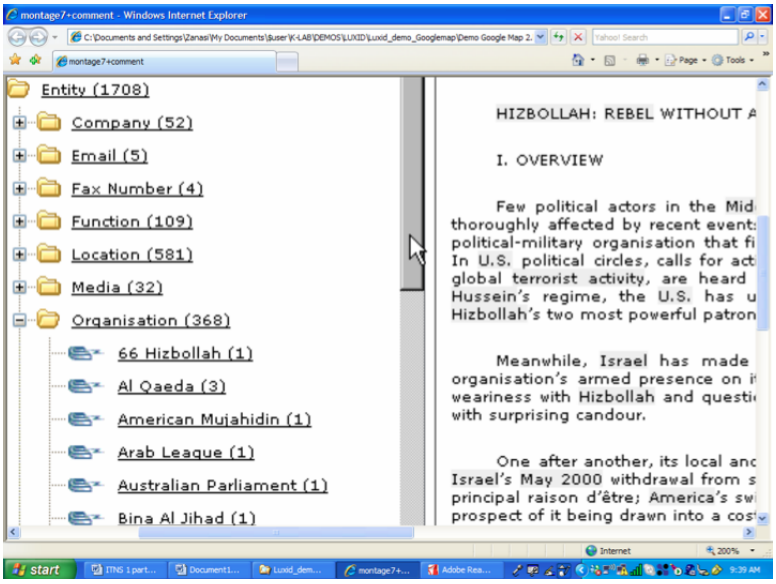

Figura 4 – Tela mostrando na esquerda palavras mais monitoradas e na direita documentos

É possível também analisar as pessoas que se envolvem de alguma forma com terroristas, através das comunidades virtuais, redes sociais e fóruns. Além disso, é possível detectar similaridades entre discursos de pessoas e organizações aparentemente não ligadas entre si, o que poderia apontar para facções ou alianças ocultas.

3. MATERIAL E MÉTODO

Na seção quatro (Resultados) será apresentada, passo a passo, uma aplicação de técnicas de *Text Mining* usando como base de dados pequenos textos extraídos da rede social *Twitter*. As subseções abaixo apresentam o *software* utilizado, como foi feita a extração dos dados e que técnicas de *Text Mining* foram aplicadas.

3.1 *Software* R

O *software* escolhido para aplicar a descoberta de conhecimento em textos neste estudo foi o R (ou R-*project*), que pode ser obtido gratuitamente no *site* http://www.r-project.org. O R é um *software* livre e amplamente usado na área de estatística e *data mining*. Possui um conjunto de funções pré-definidas, mas outras funções podem ser instaladas através de pacotes que estão disponíveis para *download* no mesmo *site* supra citado, ou através do próprio programa.

Os pacotes mais usados neste trabalho foram: *tm*, *twitteR*, *ROAuth* e *wordcloud*. Sendo o *tm*, o pacote que contém as principais funções de *Text Mining*; *twitteR* e *ROAuth*, os que permitem a conexão com a rede social para realizar a extração dos dados; e o *wordcloud*, que permite a construção de uma nuvem de palavras com os termos mais frequentes.

3.2 Dados do *Twitter*

Os dados usados neste estudo são provenientes do *Twitter*, rede social que se caracteriza por pequenos textos de até 140 caracteres.

Utilizando o R, através do comando *usertimeline* do pacote *twitteR*, foi possível coletar os *tweets* (pequenos textos) dos usuários que fizeram parte do estudo. Mas, para o R conectar com a API do *Twitter*, é preciso realizar uma autenticação, seguindo as instruções da seção 3 do documento de apoio encontrado no endereço http://cran.r-project.org/web/packages/twitteR/vignettes/twitteR.pdf.

Após conectado com a API do *Twitter*, o comando *usertimeline* foi usado da seguinte forma, abaixo representada, onde a variável *Tweets* recebe os textos mais recentes do usuário selecionado.

```
Tweets <- userTimeline("zerohora", n=3000)
```

Nesse caso, a variável recebeu os 3000 *tweets* mais recentes a partir do *Twitter* da Zero Hora, jornal de Porto Alegre. A extração foi feita às 19 horas do dia 19 de Junho de 2013. O número 3000 não foi escolhido por acaso, percebeu-se que a API do *Twitter* não respondeu de forma satisfatória para consultas acima de 3200 *tweets* usando o comando *usertimeline*.

3.3 Técnicas de *Text Mining*

Segundo Flynn (2010), o *Text Mining* pode ser dividido em duas grandes fases: 1) pré-processamento e integração dos dados não-estruturados, e 2) análise estatística dos dados pré-processados para extrair conteúdo do texto. A Figura 5 ilustra as fases da aplicação apresentada na seção 4 deste trabalho.

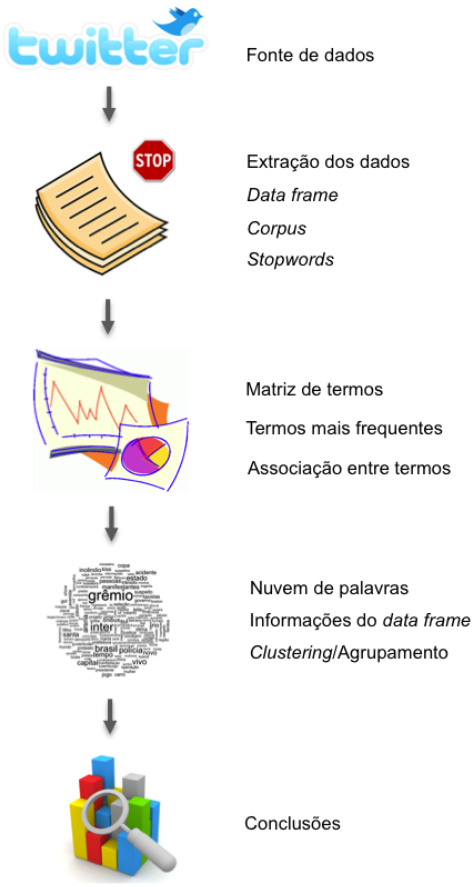

Figura 5 – Etapas da aplicação de *Text Mining* com os dados do *Twitter*

Como descrito na subseção anterior, os dados foram obtidos do *Twitter* da Zero Hora. A partir do *data frame* criado com os *tweets*, é extraída a coluna que contém os textos publicados, formando assim o *corpus* – coleção dos textos – e posteriormente as suas outras colunas também são utilizadas, pois contém

informações, como por exemplo, data de publicação dos *tweets*. No *corpus* as letras maiúsculas são transformadas em minúsculas e retiram-se as pontuações, números, endereços de sites e também *stopwords* – palavras de parada.

Após isso, é criada uma matriz de termos para identificar os termos mais frequentes e possíveis associações entre eles. Além disso, a técnica de *Clustering* é aplicada para identificar expressões compostas nos *tweets* e também agrupá-los conforme seus assuntos. Para mostrar os termos mais frequentes de forma visual, é construída uma nuvem de palavras.

Com base nas informações resultantes de cada etapa, são feitas conclusões e interpretações a respeito dos dados analisados.

4. RESULTADOS

Nesta seção, são mostradas, passo a passo, as fases básicas da mineração de texto em uma aplicação prática. Na fase de pré-processamento, utilizou-se os dados obtidos na subseção 3.2 Dados do *Twitter*.

4.1 *Data frame*

O comando abaixo transforma o objeto *Tweets* em um *data frame*. A Figura 6 mostra um registro do *data frame* df.

```
df <- do.call("rbind", lapply(Tweets, as.data.frame))
```

text	Haddad e Alckmin voltam atrás e reduzem preço da passagem de ônibus em São Paulo http://t.co/B7mkX0VwjH
favorited	FALSE
replyToSN	<NA>
created	19/06/13
truncated	FALSE
replyToSID	<NA>
id	347463509929508000
replyToUID	<NA>
statusSource	HootSuite
screenName	zerohora
retweetCount	26
retweeted	FALSE

Figura 6 – Um registro do *data frame* df (*tweet* do Twitter da Zero Hora)

4.2 *Corpus*

Para extrair a parte textual e construir o *corpus* – coleção de textos – foi usado o comando abaixo, onde ZHCorpus recebe, então, a coleção de *tweets* extraídos do *Twitter* da Zero Hora.

```
ZHCorpus <- Corpus(VectorSource(df$text))
```

Olhando para o corpus, usando o comando *inspect* para alguns *tweets*, pode-se comparar ao que mostra a página do *Twitter* da Zero Hora, como na Figura 7.

```
inspect(ZHCorpus[4:7])
```

```
[[1]] Haddad e Alckmin voltam atrás e reduzem preço da passagem de ônibus em São Paulo http://t.co/B7mkX0VwjH
[[2]] Com show de Neymar, Brasil vence o México por 2 a 0 no Castelão. http://t.co/THsacdn3Ay
[[3]] Gol do Brasil! Jô amplia. Brasil 2 x 0 México. Siga ao vivo: http://t.co/rGP0prkake
[[4]] Papa deixa que um fã de Messi suba no papamóvel.  http://t.co/Tx2BFvmiXX
```

Figura 7 – *Tweets* extraídos do Twitter da Zero Hora

De posse do *corpus*, as letras maiúsculas foram transformadas em minúsculas, e procedeu-se também a retirada das pontuações, números e endereços de sites, utilizando os seguintes comandos.

```
ZHCorpus <- tm_map(ZHCorpus, tolower)
```

```
ZHCorpus <- tm_map(ZHCorpus, removePunctuation)
ZHCorpus <- tm_map(ZHCorpus, removeNumbers)
removeURL <- function(x) gsub("http[[:alnum:]]*", "", x)
ZHCorpus <- tm_map(ZHCorpus, removeURL)
```

Após aplicar os comandos, os *tweets* ficaram como mostram os exemplos abaixo.

```
[[1]]  haddad e alckmin voltam atrás e reduzem preço da passagem de ônibus em são paulo
[[2]]  com show de neymar brasil vence o méxico por  a  no castelão
[[3]]  gol do brasil jô amplia brasil  x  méxico siga ao vivo
[[4]]  papa deixa que um fã de messi suba no papamóvel
```

4.3 *Stopwords*

Os comandos abaixo foram usados para remover as *stopwords*, que são palavras não desejadas na contagem dos termos mais frequentes por não conterem significado relevante, como preposições, conjunções, artigos e outros termos específicos. O *software* R possui listas de *stopwords* para algumas línguas.

A lista de palavras do Inglês é bem completa, enquanto a do Português é pequena e insuficiente. Neste estudo foi usada uma lista de *stopwords* encontrada no blog do professor Stanley Loh (*http://miningtext.blogspot.com.br/2008/11/listas-de-stopwords-stoplist-portugues.html*), como mostra o comando abaixo.

```
ZHStopwords <- c('a', 'à', 'agora', 'ainda', 'alguém', 'algum', 'alguma', 'algumas', 'alguns', 'ampla', 'amplas', 'amplo', 'amplos', 'ante', 'antes', 'ao', 'aos', 'após', 'aquela', 'aquelas', 'aquele', 'aqueles', 'aquilo', 'as', 'até', 'através', 'cada', 'coisa', 'coisas', 'com', 'como', 'contra', 'contudo', 'da', 'daquele', 'daqueles', 'das', 'de', 'dela', 'delas', 'dele', 'deles', 'depois', 'dessa', 'dessas', 'desse', 'desses', 'desta', 'destas', 'deste', 'deste', 'destes', 'deve', 'devem', 'devendo', 'dever', 'deverá', 'deverão', 'deveria', 'deveriam', 'devia', 'deviam', 'disse', 'disso', 'disto', 'dito', 'diz', 'dizem', 'do', 'dos', 'e', 'é', 'ei', 'ela', 'elas', 'ele', 'eles', 'em', 'enquanto', 'entre', 'era', 'essa', 'essas', 'esse', 'esses', 'esta', 'está', 'estamos', 'estão', 'estas', 'estava', 'estavam', 'estávamos', 'este', 'estes', 'estou', 'eu', 'fazendo', 'fazer', 'feita', 'feitas', 'feito', 'feitos', 'foi', 'for', 'foram', 'fosse', 'fossem', 'grande', 'grandes', 'há', 'isso', 'isto', 'já', 'la', 'lá', 'lhe', 'lhes', 'lo', 'mas', 'me', 'mesma', 'mesmas', 'mesmo', 'mesmos', 'meu', 'meus', 'minha', 'minhas', 'muita', 'muitas', 'muito', 'muitos', 'na', 'não', 'nas', 'né', 'nem', 'nenhum', 'nessa', 'nessas', 'nesta', 'nestas', 'ninguém', 'no', 'nos', 'nós', 'nossa', 'nossas', 'nosso', 'nossos', 'num', 'numa', 'nunca', 'o', 'os', 'ou', 'outra', 'outras', 'outro', 'outros', 'para', 'pela', 'pelas', 'pelo', 'pelos', 'pequena', 'pequenas', 'pequeno', 'pequenos', 'per', 'perante', 'pode', 'pôde', 'podendo', 'poder', 'poderia', 'poderiam', 'podia', 'podiam', 'pois', 'por', 'porém', 'porque', 'posso', 'pouca', 'poucas', 'pouco', 'poucos', 'primeiro', 'primeiros', 'própria', 'próprias', 'próprio', 'próprios', 'quais', 'qual', 'quando', 'quanto', 'quantos', 'que', 'quem', 'são', 'se', 'seja', 'sejam', 'sem', 'sempre', 'sendo', 'será', 'serão', 'seu', 'seus', 'si', 'sido', 'só', 'sob', 'sobre', 'sua', 'suas', 'talvez', 'também',
```

'tampouco', 'te', 'tem', 'tendo', 'tenha', 'ter', 'teu', 'teus', 'ti', 'tido', 'tinha', 'tinham', 'toda', 'todas', 'todavia', 'todo', 'todos', 'tu', 'tua', 'tuas', 'tudo', 'última', 'últimas', 'último', 'últimos', 'um', 'uma', 'umas', 'uns', 'vendo', 'ver', 'vez', 'vindo', 'vir', 'vos', 'vós')

```
ZHCorpus <- tm_map(ZHCorpus, removeWords, ZHStopwords)
```

O comando acima removeu dos *tweets* da Zero Hora as palavras contidas na lista ZHStopwords. A escolha das palavras de parada, ou *stopwords*, é importante para a descoberta de conhecimento em texto, porque ao excluir uma palavra da análise, deve-se ter cuidado para não perder informação relevante. Em contrapartida, não se pode manter palavras que são vazias de significados, pois atrapalham o processo.

Após a retirada das *stopwords*, foram identificados os termos mais frequentes nos *tweets* (usando os mesmos procedimentos que serão descritos nas subseções 4.4 e 4.5) e percebeu-se que ainda restaram palavras irrelevantes nos textos. Além da remoção das palavras de parada tradicionais, é preciso verificar para o contexto da aplicação quais são seus termos específicos. Neste estudo foram identificadas as palavras abaixo como expressões muito usadas nas publicações da Zero Hora e foram então removidas.

```
ZHStopwords2 <- c('abre', 'acompanhe', 'afirma', 'ano', 'anos', 'anuncia', 'aponta', 'blog', 'boa', 'bom', 'chega', 'começa', 'confira', 'confirma', 'deixa', 'dia', 'dias', 'durante', 'fala', 'faz', 'foto', 'fotos', 'frente', 'grupo', 'hoje', 'horas', 'imagens', 'leia', 'cinco', 'mais', 'menos', 'dois', 'neste', 'número', 'um', 'obrigado', 'olá', 'ótima', 'primeira', 'primeiro', 'quartafeira', 'quase', 'quatro', 'quintafeira', 'saiba', 'duas', 'mil', 'janeiro', 'milhões', 'morre', 'vence', 'segundafeira', 'segundo', 'seis', 'semana', 'ser', 'sextafeira', 'siga', 'terçafeira', 'três', 'vai', 'veja', 'via', 'vídeo', 'você', 'volta', 'zero', 'hora', 'porto', 'alegre')

ZHCorpus <- tm_map(ZHCorpus, removeWords, ZHStopwords2)
```

4.4 Matriz de termos

Após a remoção das *stopwords*, os *tweets* foram transformados em uma matriz de termos, utilizando o comando a seguir, onde as linhas são as palavras e as colunas são os documentos (*tweets*). Os valores dessa matriz são a quantidade de vezes que cada termo aparece em cada documento.

```
ZHTdm <- TermDocumentMatrix(ZHCorpus, control = list(minWordLength = 1))
```

Olhando para um recorte da matriz de termos, para exemplificar, pode-se ver que nos *tweets* 70 a 80, o termo “dilma“ ocorre uma vez no documento 74, como

mostra a Figura 8.

	Documento										
Termo	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
dilma	0	0	0	0	1	0	0	0	0	0	0
dilúvio	0	0	0	0	0	0	0	0	0	0	0
dimensão	0	0	0	0	0	0	0	0	0	0	0
diminui	0	0	0	0	0	0	0	0	0	0	0

Figura 8 – Matriz de termos ZHTdm

4.5 Termos mais frequentes

A partir da matriz de termos, pode-se identificar os termos mais frequentes usando o comando abaixo. Neste caso o parâmetro *lowfreq* está configurado para apresentar apenas as palavras que ocorreram no mínimo 60 vezes nos documentos, como mostra a Figura 9.

```
findFreqTerms(ZHTdm, lowfreq=60)
```

arena	novo	tempo
brasil	polícia	transitozh
capital	rio	vivo
estado	santa	zheconomia
grêmio	segundocaderno	zhesportes
inter	sul	zhpolitica

Figura 9 – Termos que ocorreram no mínimo 60 vezes nos *tweets*

Este ponto do estudo é onde a primeira vez tem-se acesso a uma informação relevante para a descoberta de conhecimento. As palavras mais frequentes nas publicações da Zero Hora, que a Figura 9 apresenta, indicam os assuntos, acontecimentos e personagens mais comentados nos *tweets*, no período em que os dados foram extraídos.

Construiu-se também um gráfico, apresentado na Figura 10, para visualizar a frequência de cada termo. Os comandos abaixo geram um gráfico de barras para as palavras que ocorreram pelo menos 60 vezes nos *tweets* da Zero Hora. A função *rowSums* é usada para somar as ocorrências de cada termo nos documentos na matriz de termos ZHTdm. Para criar o gráfico foi usada a função *qplot* do pacote *ggplot2*.

```
termosFreq <- rowSums(as.matrix(ZHTdm))
```

```
termosFreq <- subset(termosFreq, termosFreq>=60)
library(ggplot2)
qplot(names(termosFreq), termosFreq, geom="bar", xlab="termos") + coord_flip()
```

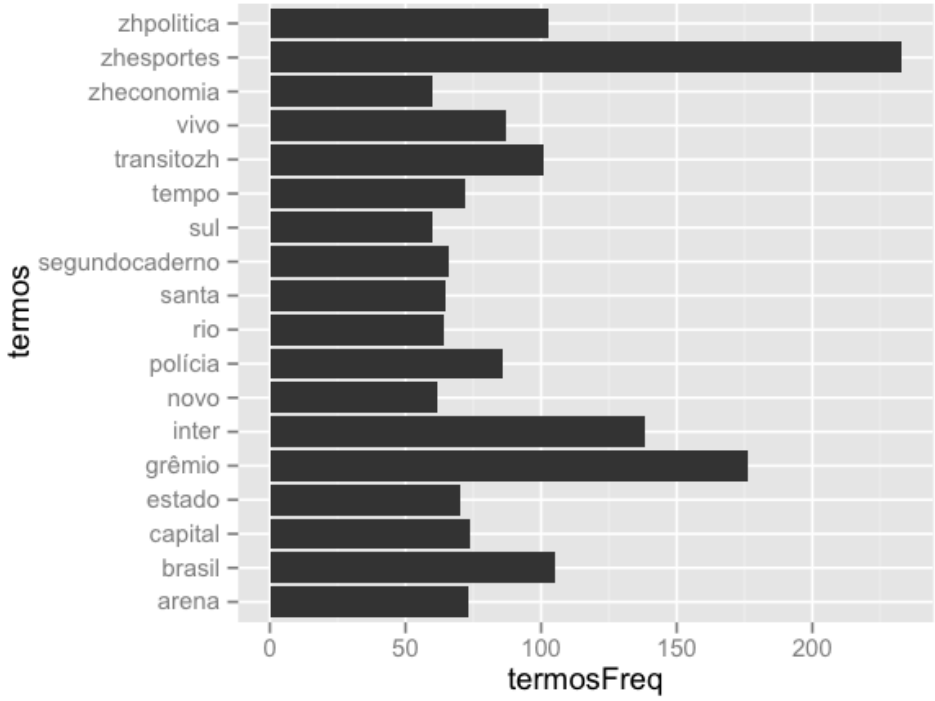

Figura 10 – Gráfico de frequência de termos

4.6 Associação entre termos

A função *findAssocs*, pertencente ao pacote *tm* do *software* R, segundo documentação dos autores, computa através da matriz de termos a correlação entre todas as palavras encontradas nos documentos analisados, e apresenta apenas as que tiverem uma correlação maior que o limite estipulado pelo usuário. Informações mais detalhadas podem ser encontradas no artigo "*Text Mining Infrastructure in R*", criado pelos autores do pacote *tm*.

Utilizando as palavras mais frequentes nos *tweets* da Zero Hora, procurou-se os termos associados a elas através dos comandos abaixo, onde os termos procurados estão entre aspas simples e o valor ao lado é a correlação mínima que se deseja encontrar entre as palavras.

```
findAssocs(ZHTdm, 'zhpolitica', 0.15)
stf      dilma   mensalão  calheiros      ccj      fux    záchia  congresso
0.27     0.21     0.19         0.17        0.17     0.17     0.17       0.16
```

O termo “zhpolitica“ retornou as palavras acima como as palavras mais associadas a ele. Percebe-se a sigla “stf“ – Superior Tribunal Federal – como o termo mais correlacionado ao assunto política nas publicações da Zero Hora, seguido da palavra “dilma“, que é a Presidente do Brasil. Logo atrás vem os termos “mensalão“ e “calheiros“, que remetem a acontecimentos no meio político do país.

Relacionado ao assunto esporte, as palavras mais associadas encontradas nos *tweets* mostram “grêmio“ e “inter“, que são os maiores times do Rio Grande do Sul (local do jornal Zero Hora), e, “tempo“ e “real“ por causa da expressão “Siga em tempo real pelo zhEsportes“ ser bastante usada.

```
findAssocs(ZHTdm, 'zhesportes', 0.15)
grêmio   real  inter  tempo  frança
  0.30   0.26   0.18   0.18   0.17
```

A Zero Hora disponibiliza, em seu *twitter*, informações sobre o trânsito em diversos pontos da região metropolitana de Porto Alegre, o que pode ser percebido nas palavras abaixo.

```
findAssocs(ZHTdm, 'transitozh', 0.15)
lento triunfoconcepa bairrocentro castelo intenso ferimentos congestionado protásio transitoarena
 0.39        0.26         0.24     0.24    0.22      0.20         0.18      0.18         0.18
```

O comando abaixo mostra a palavra “sentido“ como a mais fortemente relacionada a “bairrocentro“, pois alguns *tweets* reportam o tráfego nas ruas especificando o sentido do fluxo, neste caso, do bairro para o centro.

```
findAssocs(ZHTdm, 'bairrocentro', 0.15)
sentido   assis congestionado protásio
   0.52   0.45         0.45     0.45
```

Outras palavras também muito frequentes no twitter da Zero Hora foram “grêmio“ e “inter“, mencionadas a pouco como termos correlacionados ao assunto esporte. Para “grêmio“, aparecem “elano“ e “vargas“, jogadores que costumam se destacar nos jogos, e “luxemburgo“ que é o técnico do time. Relacionado ao termo “inter“ foram identificadas as palavras “juventude“ e “criciúma”, times que foram enfrentados, e, “forlán“, jogador estrangeiro do Inter.

```
findAssocs(ZHTdm, 'grêmio', 0.15)
```

zhesportes	elano	zhgremio	placar	vargas	arena	libertadores	luxemburgo	gol
0.3	0.23	0.20	0.19	0.18	0.17	0.17	0.16	0.15

```
findAssocs(ZHTdm, 'inter', 0.15)
```

juventude	gol	criciúma	marca	taça	zhinter	forlán	placar	zhesportes
0.31	0.28	0.21	0.21	0.21	0.21	0.20	0.19	0.18

Em todo o Rio Grande do Sul, mais especificamente em Porto Alegre, existe uma rivalidade muito grande entre as torcidas do Inter e do Grêmio. Nesse contexto, o estudo aponta para um fato curioso, quando mostra a palavra “gol“ associada aos dois times em questão, só que em grandezas diferentes. A correlação do termo “gol“ com o Grêmio é de 0.15, enquanto com o Inter, é de 0.28, quase o dobro. Não se está dizendo que um time faz mais gols que o outro, mas a análise mostra que, no período estudado, existem mais *tweets* com as palavras “gol“ e “inter“ do que “gol“ e “grêmio“. Em contrapartida, o termo “grêmio“ é mais frequente que o termo “inter“.

Apesar do jornal Zero Hora também publicar notícias do mundo todo, o seu foco é em notícias nacionais, então, as notícias em geral não usam o termo “brasil“ para identificar o local dos acontecimentos. O termo em questão aparece fortemente relacionado ao esporte, em *tweets* que cobrem notícias sobre a seleção brasileira de futebol, como mostram as associações abaixo.

```
findAssocs(ZHTdm, 'brasil', 0.15)
```

França	goool	bolívia	hernanes	amistoso	goleada	lucas	assis	copa	neymar
0.42	0.41	0.36	0.26	0.22	0.19	0.19	0.17	0.17	0.17

A análise dos termos mais frequentes e das suas associações, apesar de ser uma ideia simples, traz muita informação relevante e até novos conhecimentos, porém, demanda um certo envolvimento e conhecimento prévio do usuário. Mas, ainda assim, é muito eficaz para extrair conhecimento de um base de dados não-estruturada como o texto.

4.7 Nuvem de palavras

Com os termos mais frequentes, foi construída uma nuvem de palavras – gráfico que apresenta as palavras com tamanhos diferentes, de acordo com seu número de ocorrências, quanto mais frequente, maior é a fonte da palavra. Mas para isso, primeiro procedeu-se a remoção de alguns termos específicos, como mostram

os comandos abaixo, por serem palavras que indicam outras contas de *twitter* da Zero Hora ou *hashtags* – palavras que identificam o assunto do *tweet* e que funcionam como *links*.

```
ZHStopwords3 <- c('zhanos', 'zheconomia', 'zhesportes', 'zhgremio', 'zhmundo', 'zhpolitica', 'opiniaozh', 'segundocaderno', 'transitozh', 'aovivozh', 'capa')

ZHCorpus <- tm_map(ZHCorpus, removeWords, ZHStopwords3)
```

Com o novo *corpus*, agora sem as palavras acima, novamente foi criada a matriz de termos para ser usada na construção da nuvem de palavras, utilizando o seguinte comando.

```
ZHTdm <- TermDocumentMatrix(ZHCorpus, control = list(minWordLength = 1))
```

Para criar a nuvem de palavras é necessário a instalação do pacote *wordcloud* e seus pacotes dependentes. Os comandos abaixo foram usados para carregar os pacotes necessários e construir a nuvem usando a nova matriz de termos.

```
library(wordcloud)
m <- as.matrix(ZHTdm)
v <- sort(rowSums(m), decreasing=TRUE)
myNames <- names(v)
d <- data.frame(word=myNames, freq=v)
wordcloud(d$word, d$freq, min.freq=16)
```

O parâmetro *min_freq* do comando *wordcloud* foi configurado para a nuvem de palavras ser feita somente com os termos que ocorrem pelo menos 16 vezes, para que o gráfico ficasse legível e não poluído visualmente. A Figura 11 mostra a nuvem de palavras resultante.

A nuvem de palavras do *twitter* da Zero Hora mostra em destaque os termos “grêmio“, “inter“, “brasil“, “polícia“ e outros. Este tipo de gráfico é uma boa opção por mostrar de forma visual o que mais é publicado nos *tweets*. Como a Zero Hora é um jornal, o seu *twitter* publica micro notícias, o que já é uma espécie de resumo da informação. E através do *text mining* foi possível resumir ainda mais os acontecimentos e assuntos mais comentados a partir dos *tweets*.

Assim, uma nuvem de palavras como essa, resume em uma imagem simples os assuntos do momento em que os dados foram extraídos. Pode-se criar análises como essa em tempo real, mostrando o que está acontecendo em um determinado local.

No site *http://www.causabrasil.com.br* pode ser encontrada uma espécie de nuvem de palavras, que também usa técnicas de *text mining* para extrair informação de redes sociais como *Facebook*, *Twitter*, *Youtube* e outras. Neste caso, a fonte de dados usada são textos sobre os protestos que estão acontecendo no Brasil. O diferencial é que a nuvem está sendo alimentada em tempo real e o gráfico resume a informação de várias fontes.

Figura 11 – Nuvem de palavras do *twitter* da Zero Hora

4.8 Informação do *data frame*

Ainda com relação ao estudo do *twitter* da Zero Hora, também foram analisadas as datas de publicação dos *tweets*, suas origens e também quantos *retweets* receberam – quando um *tweet* é publicado novamente por outra pessoa a partir do *twitter* do autor.

Os comandos abaixo retornaram a origem dos *tweets* da Zero Hora, ou seja, a partir de qual serviço foram publicados. Através do *data frame* df, na coluna *statusSource* esta informação é extraída.

```
sources <- df$statusSource
sources <- gsub("</a>", "", sources)
sources <- strsplit(sources, ">")
sources <- sapply(sources, function(x) ifelse(length(x) > 1, x[2], x[1]))
table(sources)
```

Como mostra a Figura 12, o serviço mais usado pela Zero Hora na publicação dos seus *tweets* é o *HootSuite*, sistema de gerenciamento e controle de mídias sociais via *web*.

Origem	Tweets	%
HootSuite	2976	99,2%
Storify	14	0,5%
Bitly	5	0,2%
Web	3	0,1%
Pinterest	2	0,1%

Figura 12 – Tabela de frequência da origem dos *tweets* da Zero Hora

A partir das colunas *created* e *retweetCount* do *data frame* df, foram extraídas as datas dos *tweets* e a contagem de *retweets* de cada publicação, utilizando o código abaixo.

```
retweets <- df$retweetCount
tweetsdatas <- df$created
write.csv(retweets, file='cont_retweets.csv')
write.csv(tweetsdatas, file='tweets_datas.csv')
```

Os dados foram convertidos em arquivos de extensão *csv* e após isso, manipulados com a ajuda do *software* Excel. Identificou-se que os 3 mil *tweets* usados neste estudo compreendem o período de 3 de abril de 2013 a 19 de junho de 2013, o que equivale a 76 dias. Baseado nesta amostra, a Zero Hora publicou em média 39 *tweets* por dia e a Figura 13 mostra a distribuição das publicações por hora. Há uma concentração maior de *tweets* entre às 11 da manhã e meio dia, horário em que a maioria das pessoas está acordada e em atividade.

Os *tweets* da Zero Hora receberam em média 11 *retweets*, ou seja, cada *tweet* foi publicado novamente 11 vezes em média por outros usuários da rede social. A publicação que recebeu mais *retweets* no período analisado, foi publicada

novamente 247 vezes. O texto deste *tweet* foi “Inter é tricampeão gaúcho!“, publicado no dia 5 de maio de 2013.

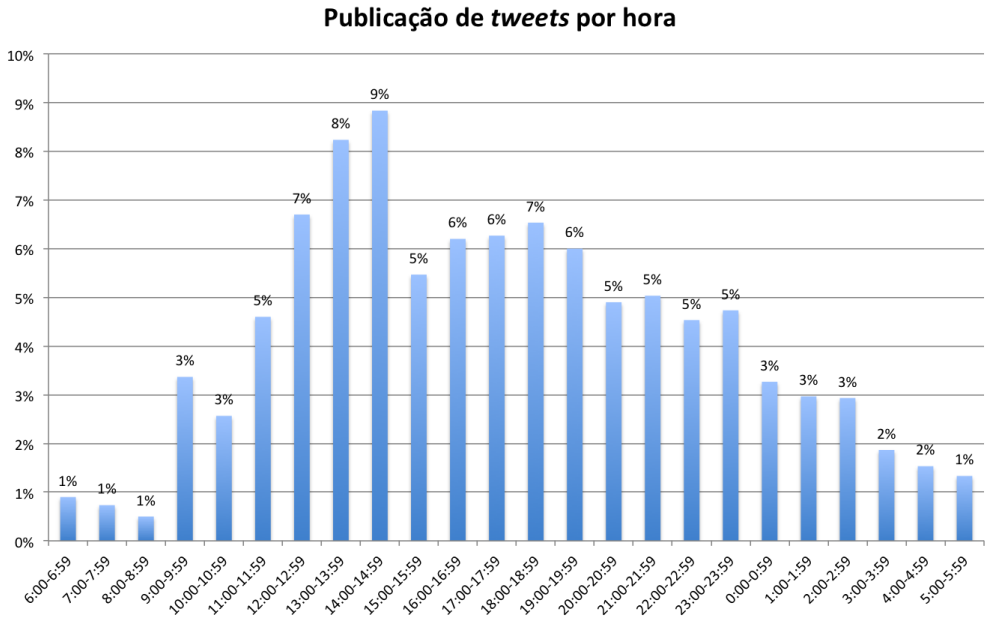

Figura 13 – Distribuição da publicação de *tweets* da Zero Hora

4.9 *Clustering*

Para encontrar *clusters* ou agrupamentos de palavras nos *tweets* da Zero Hora, aplicou-se o código abaixo à matriz de termos ZHTdm. Foram removidos os termos mais esparsos usando o parâmetro *sparse* (são usados na análise apenas os termos mais frequentes, com menos zeros na matriz), e a matriz resultante ZHTdm2 ficou com 65 palavras.

Após as distâncias entre termos serem calculadas pela função *dist*, usando o método *ward*, o *clustering* hierárquico foi processado para os *tweets* e a Figura 14 mostra o resultado.

```
ZHTdm2 <- removeSparseTerms(ZHTdm, sparse=0.992)
m2 <- as.matrix(ZHTdm2)
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="ward")
plot(fit)
```

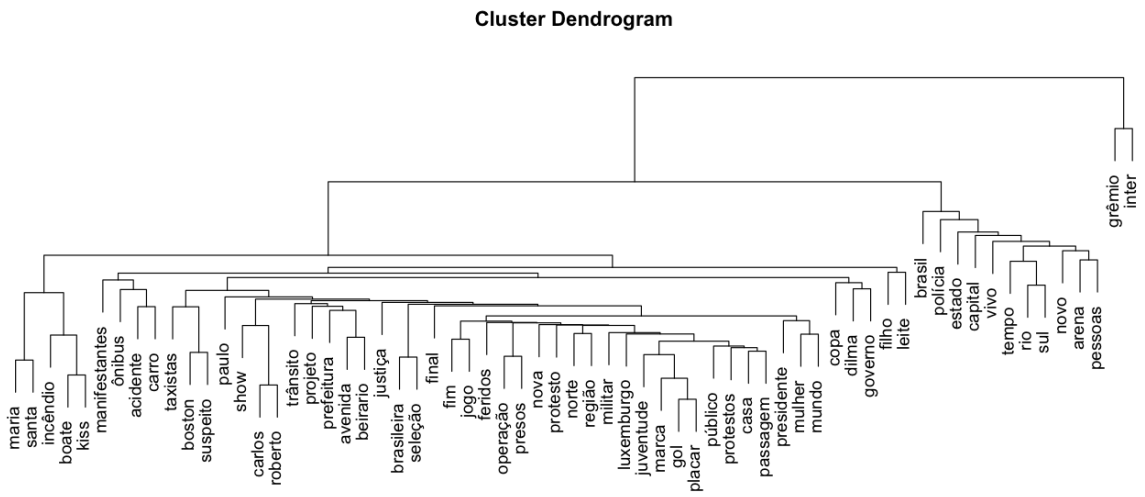

Figura 14 – *Clustering* dos *tweets* da Zero Hora

O primeiro cluster é o que contém os termos “grêmio“ e “inter“, que são as palavras mais frequentes deste estudo e os dois maiores times do Rio Grande do Sul, como dito anteriormente. Outro agrupamento que chama a atenção é o primeiro da esquerda, por se tratar do incêndio acontecido na boate Kiss na cidade de Santa Maria, caso muito noticiado que chocou o Brasil e o mundo. Na nuvem de palavras os termos desse *cluster* já tinham aparecido como mais frequentes, porém separados como palavras independentes, e aqui eles aparecem juntos.

Através do *clustering* dos *tweets*, também foi possível identificar expressões compostas como show do Roberto Carlos e seleção brasileira, e grupos de palavras como “dilma“, “copa“ e “governo“.

Outra técnica de *clustering* também foi aplicada aos *tweets* para agrupá-los de acordo com seus assuntos, usando o algoritmo *k-means*, como mostram os comandos abaixo. Aqui, a matriz de termos ZHTdm2 foi transposta e os *tweets* foram agrupados de acordo com o número de *clusters* escolhido e configurado no parâmetro k.

```
m3 <- t(m2)
k <- 7
kmeansResult <- kmeans(m3, k)
for (i in 1:k) {
        cat(paste("cluster ", i, ": ", sep=""))
        s <- sort(kmeansResult$centers[i,], decreasing=T)
        cat(names(s)[1:4], "\n")
        print(head(Tweets[which(kmeansResult$cluster==i)]))
}
```

Abaixo os 7 *clusters* que resultaram, com as palavras que os identificam e uma amostra de *tweets* de cada grupo.

cluster 1: dilma governo presidente projeto
"Aprovação do governo Dilma caiu para 55%, segundo pesquisa CNI/Ibope."
"Governo avalia conduta da BM como correta, mas investigará excessos."
"Dilma: as vozes da rua devem ser ouvidas. (via @zhpolitica)"
"Dilma: manifestações são legítimas e próprias da democracia"
"Dilma confirma presença na final da Copa das Confederações, apesar das vaias na abertura"

cluster 2: polícia rio sul suspeito
"Direto do Rio de Janeiro, David Coimbra e Eduardo Gabardo debatem antes do jogo da Seleção Brasileira"
"#protestosp RT @rlopesreporter: Polícia usa bombas de efeito moral na frente do Palácio dos Bandeirantes"
"Pelo menos quatro pessoas ficam feridas em confronto com a polícia durante protesto em Porto Alegre"
"RT @aovivozh: Manifestantes tentam avançar em direção à Avenida Erico Verissimo, mas a polícia usa novas bombas de efeito moral."
"RT @aovivozh: Movimento reúne milhares de manifestantes no Rio de Janeiro e em Belo Horizonte"

cluster 3: roberto carlos show arena
"Roberto Carlos tenta impedir venda de livro sobre a Jovem Guarda."
"50 mil pessoas assistiram Roberto Carlos na Arena. Saiba todos os detalhes"
"Perdeu o show do Roberto Carlos na Arena? Confira trechos da apresentação em vídeo"
"Galeria de fotos: veja como foi o show de Roberto Carlos na Capital"
"Roberto Carlos ganha bolo de aniversário e chuva de fogos de artifício em show na Arena do Grêmio em Porto Alegre"

cluster 4: inter brasil capital estado
"@eifrodo Sim, estamos com problemas técnicos. Te avisamos quando estiver ok."
"RT @zh_esportes: Inter deve confirmar venda de Fred ao Shakhtar"
"Haddad e Alckmin voltam atrás e reduzem preço da passagem de ônibus em São Paulo"
"Com show de Neymar, Brasil vence o México por 2 a 0 no Castelão"
"Gol do Brasil! Jô amplia. Brasil 2 x 0 México. Siga ao vivo"

cluster 5: grêmio arena jogo tempo
"RT @zh_esportes: Jornal italiano garante que Eduardo Vargas fica no Grêmio até o final do ano"
"RT @zh_esportes: Por apoio da torcida, Grêmio levará letreiro de campeão do mundo à Arena"
"Após meses de negociação, Fábio Koff assegura: A Arena é do Grêmio."
"RT @zh_esportes: Conselho aprova nesta segunda mudanças na parceria entre Grêmio e OAS. Veja o que irá à votação"
"RT @zh_gremio: Espero voltar ao Grêmio e jogar na Arena, diz argentino Facundo Bertoglio"

cluster 6: santa maria kiss incêndio
"Após quatro meses internada, jovem chega a Santa Maria"
"Penúltima vítima da boate Kiss internada recebe alta e já está a caminho de Santa Maria."
"Somente as famílias estão sendo condenadas, afirma pai de vítima da tragédia de Santa Maria."
"RT @aovivozh: Ao vivo: Acompanhe a divulgação da investigação do inquérito da BM sobre a tragédia em Santa Maria"

"Incêndio na Kiss: Brigada divulga IPM da tragédia de Santa Maria nesta quarta-feira. Saiba mais"

cluster 7: carro militar acidente feridos
"RT @transitozh: Após acidente, carro invade garagem de prédio e atinge mais dois veículos em Porto Alegre"
"Manifestantes terão de se apresentar à Brigada Militar em dias de protestos em Porto Alegre"
"RT @zh_esportes: FOTO: como ficou o carro do atacante Vargas após acidente no Norte do Estado."
"Corpo é encontrado em porta-malas de carro em Montenegro com dois tiros no peito"
"RT @transitozh: RT @maurosaraivajr: Acidente na Rua do Presídio com Aparício Borges. Trânsito lento. Entre carro e moto. Sem feridos graves"

Através da técnica de *clustering*, obtém-se os grupos de *tweets* como mostrado acima, e então pode-se por exemplo, fazer também uma análise mais profunda em um *cluster* específico para descobrir conhecimento a respeito de um assunto em especial.

Os *tweets* da Zero Hora são pequenas notícias e, por se tratar de um Jornal, informam acontecimentos e informações de utilidade pública. Através do agrupamento desses *tweets*, percebeu-se que várias publicações são feitas em relação ao mesmo assunto, até porque existem fatos que são discutidos dias ou até mesmo semanas. E os grupos refletiram as palavras mais frequentes encontradas nos tweets.

5. CONSIDERAÇÕES FINAIS

Assim como o *Data Mining*, o *Text Mining* dispõe de técnicas diversas que podem ser usadas em conjunto ou separadas, mas o usuário precisa definir quais são as mais aderentes para realizar a descoberta de conhecimento a partir da sua base de textos.

Neste trabalho foi apresentada a técnica de *Text Mining* de forma geral e alguns exemplos de aplicações. Além disso, foi mostrado um estudo feito com o *Twitter* do jornal Zero Hora. O crescimento do uso de redes sociais aumenta a necessidade de se explorar esse tipo de informação.

O *Twitter*, rede social usada como fonte de dados neste trabalho, por sua característica de permitir apenas postagens com no máximo 140 caracteres, já é por si só um local onde as informações são resumidas. Como o perfil escolhido para realizar o estudo foi o do jornal Zero Hora, cada *tweet* representa informações ou manchetes de notícias. Então, ao analisar os *tweets* de um jornal usando as técnicas de *Text Mining*, pode-se obter um resumo de fatos e acontecimentos.

Com os termos mais frequentes encontrados nos textos publicados, foi construída uma nuvem de palavras. Apesar da ideia ser simples e de fácil execução,

esta forma visual foi muito eficaz para transmitir informações sobre as publicações do jornal no período avaliado. Ao olhar para a nuvem, é fácil perceber que predominam, nas postagens da Zero Hora, informações sobre os times Grêmio e Inter. A grande quantidade de publicações sobre o assunto esporte, pode ser devido ao público que o jornal quer alcançar através do seu *Twitter*.

Apesar da nuvem de palavras resumir de forma visual grande quantidade de textos ao mostrar os termos mais usados, a técnica mostrou-se ineficaz para as expressões compostas por mais de uma palavra. Em contrapartida, o *Clustering* Hierárquico mostrou na forma de dendograma a informação que faltava na nuvem. As palavras “incêndio“, “kiss“ e “santa“, que podem ser vistas facilmente na nuvem, são um exemplo disso. Separadas podem ser interpretadas de forma equivocada, mas através do dendograma, pode-se entender que ”incêndio” está relacionado com “boate kiss“ e “santa maria“.

Para complementar a análise, o agrupamento dos *tweets* dividiu-os de acordo com os seus assuntos. Um dos grupos, o cluster 6, foi identificado pelas palavras “santa“, “maria“, “kiss“ e “incêndio“. Lendo manualmente alguns *tweets* do grupo mencionado, pode-se facilmente descobrir que ocorreu um incêndio na boate Kiss na cidade de Santa Maria. Percebeu-se assim que as técnicas de nuvem de palavras e *Clustering* são complementares, e mais eficazes quando são usadas juntas.

Neste trabalho, conclui-se que é viável extrair conhecimento a partir de bases textuais em redes sociais de forma simples e com baixo custo. O *software* utilizado é gratuito e o *Twitter* disponibiliza sua API sem custo algum. Além disso, existem diversos manuais sobre o pacote de *Text Mining* do R.

Tanto para obter um conhecimento mais superficial, quanto para encontrar respostas a questões específicas, o *Text Mining* apresenta-se como um conjunto de técnicas, muito eficaz para extrair informação de textos. Porém, percebeu-se que a intervenção humana ainda se faz muito necessária, os processos ainda estão longe de ser totalmente automatizados. A começar pelo fato de que existem muitas línguas, e cada uma tem suas caraterísticas específicas.

Além disso, cada tipo de texto tem seu estilo de escrita, o que faz toda a diferença no momento de retirar as palavras de parada (*stopwords*), por exemplo. A fase da retirada das *stopwords* dos textos é muito importante. A definição de quais palavras devem ser removidas faz toda a diferença nas análises posteriores. Essa questão pode ser uma sugestão de trabalhos futuros, para entender as nuances de cada contexto, e encontrar uma forma de criar listas de *stopwords* que se adaptam aos diferentes assuntos e estilos de escrita.

O crescente uso das redes sociais e da internet como um todo, faz com que muitos dados em formato de texto trafeguem pelo mundo. E as organizações que souberem como obter e entender toda essa informação, com certeza irão ter em suas mãos, conhecimento valioso para subsidiar suas tomadas de decisão.

6. REFERÊNCIAS BIBLIOGRÁFICAS

DIXON, M. *An Overview of Document Mining Technology*. 1997.

DODDS, P. S.; HARRIS, K. D.; KLOUMANN, I. M.; BLISS, C. A.; DANFORTH, C. M. *Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter*. University of Vermont, 2011.

FEINERER, I.; HORNIK, K.; MEYER, D. *Text Mining Infrastructure in R*. Journal of Statistical Software, 2008.

FELDMAN, R.; DAGAN, I. *Knowledge Discovery in Textual Databases (KDT*). Bar-Ilan University, Ramat-Gan, Israel, 1995.

FLYNN, M.; FRANCIS, L. *Text Mining Handbook*. Casualty Actuarial Society E-Forum, 2010.

SILVA, E. M. *Descoberta de Conhecimento com o uso de Text Mining: Cruzando o Abismo de Moore*. Dissertação de Mestrado (Gestão do Conhecimento e da Tecnologia da Informação) – Universidade Católica de Brasília, Brasília, 2002.

TAN, A. H. *Text mining: the state of the art and the challenges*. Kent Ridge Digital Labs, 1999.

WIVES, L. K. *Tecnologias de descoberta de conhecimento em textos aplicadas à Inteligência Competitiva*. Exame de qualificação – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

ZANASI, A. *Virtual Weapons for Real Wars: Text Mining for National Security*. CISIS, 2008.