UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

Afonso Valau de Lima Junior

Metric for seleting the number of topics in the LDA Model

Porto Alegre, 2020

Afonso Valau de Lima Junior

Metric for seleting the number of topics in the LDA Model

Dissertation for Doctoral degree in Business Administration at the School of Administration of Federal University of Rio Grande do Sul.
Supervisor: João Luiz Becker, PhD

Porto Alegre, 2020

CIP - Catalogação na Publicação

Lima Junior, Afonso Valau de
Metric for seleting the number of topics in the LDA Model / Afonso Valau de Lima Junior. -- 2020.
92 f.
Orientador: João Luiz Becker.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Escola de Administração, Programa de Pós-Graduação em Administração, Porto Alegre, BR-RS, 2020.

1. Latent Dirichlet Allocation. 2. Topic model. 3. Text analytics. 4. Operational Research. I. Becker, João Luiz, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Afonso Valau de Lima Junior

Metric for seleting the number of topics in the LDA Model

Dissertation for Doctoral degree in Business Administration at the School of Administration of Federal University of Rio Grande do Sul.

Approved research. Porto Alegre, December 23, 2020:

João Luiz Becker, PhD
Dissertation Advisor

Adriano Mendonça Souza, PhD
Committee Member

Carla Bonato Marcolin, PhD
Committee Member

Luciano Ferreira, PhD
Committee Member

Porto Alegre, 2020

Agradecimentos

Em primeiro lugar gostaria de agradecer a mim, não foi uma trajetória fácil, mas aprendi muito, cresci, superei meus limites e acredito que valeu muito, *"Caraca você conseguiu"*.

Mo dupẹ lọ awọn orixás mi ati awn nkan, igbagbọ mi, ipilẹ mi.

Agradeço minha mãe Tânia, meus irmãos Flávia e Rafael. A minha vó Solema que foi uma grande incentivadora dos meus estudos e ao meu avô Aneci. Também ao Jonatan, são 7 anos juntos, mas nos 2 últimos anos se fez presente, ajudando, dividindo sorrisos e tristezas. Minha psicóloga Denise, por ouvir minhas preocupações. Minha amiga do mestrado, Viviane. É claro que não poderia esquecer do BB 😸, Anastácia 😸 e Saimon 😸.

Agradeço ao meu orientador, prof. Dr. João Luiz Becker, pela sua dedicação, sabedoria, paciência, maestria e principalmente agradeço por acreditar no meu potencial. Meu agradecimento especial a banca examinadora: Prof. Dr. Adriano Mendonça Souza, Profa. Dra. Carla Bonato Marcolin e Prof. Dr. Luciano Ferreira, obrigado pelos comentários e sugestões.

Ao Programa de Pós-Graduação em Administração e a todos os professores do doutorado que de alguma forma contribuíram para a minha formação e também ao secretário do PPGA - Thiago Antunes da Silva Cardoso. Aos colegas de Pesquisa Operacional que sempre me ajudaram, Raphael e Alfredo.

A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pelo auxílio financeiro.

If you can dream it, you can do it.
Walt Disney.

Abstract

The latest technological trends are driving a vast and growing amount of textual data. Topic modeling is a useful tool for extracting information from large corpora of text. A topic template is based on a corpus of documents, discovers the topics that permeate the corpus and assigns documents to those topics. The Latent Dirichlet Allocation (LDA) model is the main, or most popular, of the probabilistic topic models. The LDA model is conditioned by three parameters: two Dirichlet hyperparameters (α and β) and the number of topics (K). Determining the parameter K is extremely important and not extensively explored in the literature, mainly due to the intensive computation and long processing time. Most topic modeling methods implicitly assume that the number of topics is known in advance, thus considering it demands an exogenous parameter. That is annoying, leaving the technique prone to subjectivities. The quality of insights offered by LDA is quite sensitive to the value of the parameter K, and perhaps an excess of subjectivity in its choice might influence the confidence managers put on the techniques results, thus undermining its usage by firms. This dissertation's main objective is to develop a metric to identify the ideal value for the parameter K of the LDA model that allows an adequate representation of the corpus and within a tolerable elapsed time of the process. We apply the proposed metric alongside existing metrics to two datasets. Experiments show that the proposed method selects a number of topics similar to that of other metrics, but with better performance in terms of processing time. Although each metric has its own method for determining the number of topics, some results are similar for the same database, as evidenced in the study. Our metric is superior when considering the processing time. Experiments show this method is effective.

Keywords: Latent Dirichlet Allocation. Topic model. Text analytics. Operational Research.

Resumo

As tendências tecnológicas mais recentes impulsionam uma vasta e crescente quantidade de dados textuais. Modelagem de tópicos é uma ferramenta útil para extrair informações relevantes de grandes corpora de texto. Um modelo de tópico é baseado em um corpus de documentos, descobre os tópicos que permeiam o corpus e atribui documentos a esses tópicos. O modelo de Alocação de Dirichlet Latente (LDA) é o principal, ou mais popular, dos modelos de tópicos probabilísticos. O modelo LDA é condicionado por três parâmetros: os hiperparâmetros de Dirichlet (α and β) e o número de tópicos (K). A determinação do parâmetro K é extremamente importante e pouco explorada na literatura, principalmente devido à computação intensiva e ao longo tempo de processamento. A maioria dos métodos de modelagem de tópicos assume implicitamente que o número de tópicos é conhecido com antecedência, portanto, considerando que exige um parâmetro exógeno. Isso é um tanto complicado para o pesquisador pois acaba acrescentando à técnica uma subjetividade. A qualidade dos insights oferecidos pelo LDA é bastante sensível ao valor do parâmetro K, e pode-se argumentar que um excesso de subjetividade em sua escolha possa influenciar a confiança que os gerentes depositam nos resultados da técnica, prejudicando assim seu uso pelas empresas. O principal objetivo desta dissertação é desenvolver uma métrica para identificar o valor ideal para o parâmetro K do modelo LDA que permita uma representação adequada do corpus e dentro de um tempo de processamento tolerável. Embora cada métrica possua método próprio para determinação do número de tópicos, alguns resultados são semelhantes para a mesma base de dados, conforme evidenciado no estudo. Nossa métrica é superior ao considerar o tempo de processamento. Experimentos mostram que esse método é eficaz.

Palavras-chaves: Latent Dirichlet Allocation. Modelo de Tópico. Análise de texto. Pesquisa Operacional.

Contents

1 INTRODUCTION

With the continued growth of various sources of information, where a large amount of data is generated every minute, the need for data compression, analysis and management tools is becoming evident. According to Zgurovsky and Zaychenko (2020), the estimate for this decade is that humanity, every two years doubles the volume of data produced, something positive. Still, at the same time, the author warns of the so-called information gap, that is, the ability to process, analyze and understand these data does not go at the same speed.

Ghavami (2019) points out that besides the quantity, the variety of data is also increasing because the development of hardware and software platforms, mainly for the web network, has allowed the rapid creation of large quantities and different data types mainly of text.

According to Chui et al. (2012), radio took 38 years to reach an audience of 50 million users and, for television, more than a decade (13 years). The internet took three years to get 50 million subscribers. The data is more impressive on social media; Facebook, launched in 2004, took one year to reach 50 million users, and Twitter reached 50 million in nine months. Social networks have greatly influenced human interaction on an individual and social level, highlighting the convergence of the online and offline worlds (PROVOST; FAWCETT, 2013). Fuleky (2019) mentions that active Twitter users (around 275 million) write about 500 million tweets each day.

Kar and Dwivedi (2020) emphasize that digital transformation is maturing in companies and governments in different countries. For example, India and Singapore government platforms allow citizens to express their views on government policies, empowering citizens. Ghavami (2019) mentions the electronic medical record something already implemented in many hospitals. The electronic medical record generated during a visit to the doctor contains the historical history of the patients health care, including certificates, exam reports and prescriptions.

Based on Chen et al. (2012), computational tools are needed to organize, research and understand large amounts of data, especially as our collective knowledge continues to be digitized and stored by many types of media. Davenport (2020) emphasizes that data science is a new and popular field and is increasingly a mission-critical activity for companies and organizations, requiring various tasks and skills from professionals.

Regarding textual data, Makhabel (2017) highlight they were originally available in an unstructured form. To develop a useful representation of documents, providing high quality information, the research and design of algorithms capable of discovering patterns and trends through topic modeling have been promoted.

The main purpose of topic modeling is to discover the main themes that permeate a collection of documents, discover patterns in word usage, and connect documents that share similar patterns. Pattern discovery reflects the underlying topics that come together to form the documents (JO, 2018). Figure 1 presents an overview of the transformation of a textual content (unstructured) to a representation by topics. The original source's representation plays a key role in defining topics and identifying which ones are present in each document. This results in a clear representation of documents useful for analyzing the subjects present within them (ZHANG et al., 2020). Topic extraction models can be divided into non-probabilistic and probabilistic.

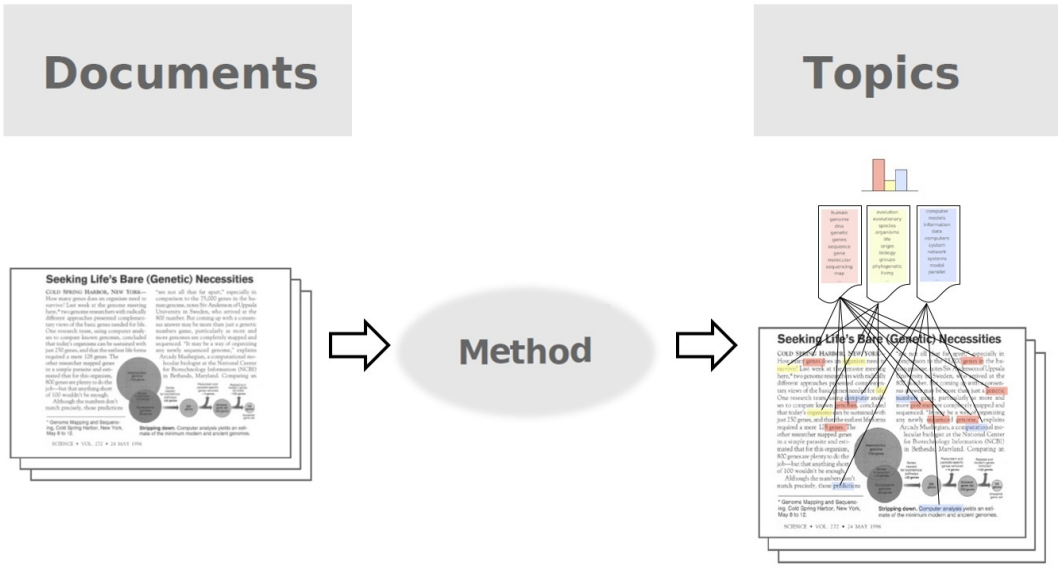


Figure 1 – Representation of documents by topics.
Source: Adapted from Blei (2012).

In non-probabilistic models, a topic can be understood as a group of terms with weights indicating the importance or significance of these terms for some subject. Discovering topics with non-probabilistic models equals grouping terms into meaningful sets (YAN et al., 2013). Among the non-probabilistic models are the Latent Semantic Analysis (LSA) model - also known as Latent Semantic Indexing (LSI) (DEERWESTER et al., 1990) and Non-Negative Matrix Factorizaton (NMF) (LEE; SEUNG, 1999). The LSA model projects both documents and terms in a lower dimension space representing the semantic concepts of the documents (AGGARWAL; ZHAI, 2012); this model aims to reduce the adverse effects caused by the synonymy and polysemy with identification of statistical associations between terms. The NMF model treats the process of topic extraction as the problem of identifying two positive semi-definite matrices such that the matrix product of them results in a good approximation of the document-term matrix. The squared Euclidean distance is usually considered as cost functions and a generalization of the Kullback-Leibler divergence (KLD) (ARORA et al., 2012). Discussions about non-probabilistic approaches are presented in Baeza-Yates et al. (1999); Manning et al. (2008); Dong et al. (2008) and Turney and Pantel (2010).

Probabilistic models have gained popularity following the introduction of the probabilistic latent semantic analysis (pLSI) model by Hofmann (1999), the first to formalize the extraction of a probabilistic topic. Although it provides a good basis for an analysis of texts, the pLSI model presents two problems. First, the topic generation process for each document is not defined, which requires determining the number of parameters that grows linearly with the number of documents and can lead to overfitting the estimated parameters. In addition, the pLSI model does not define a natural way to calculate probabilities related to a document not in the training set (BLEI et al., 2003; AGGARWAL; ZHAI, 2012; KIM et al., 2012). To avoid these problems, Blei et al. (2003) proposed the Latent Dirichlet Allocation (LDA) model, which is the main, or most popular, of the probabilistic topic models (described in more detail in section 2.1).

Numerous other extensions and applications with topic models can be found in the literature, not just for textual data; for example, Fei-Fei and Perona (2005) and Sivic et al. (2005) used topic templates to analyze images. Intense research in developing new extensions of topic models allow the incorporation of metadata and in several areas; for example, Rosen-Zvi et al. (2004) proposes the Author-Topic Model that relates topics and authorship of documents; Jo and Oh (2011) use the feeling information of words as prior to estimate the feeling of themes; Liu et al. (2007) uses correlations between numerical and textual data on product sales; Approval of candidates (LIVNE et al., 2011); Words and voting patterns (WANG et al., 2005) and Words in a social network of senders and recipients (MCCALLUM et al., 2007). Min et al. (2020), using topic modeling, explored work accident-related issues in Korea. Matsutani and Hamada (2020) use an extended Latent Dirichlet Allocation model to identify mutation signatures in cancer genomes. Sommeria-Klein et al. (2020) used Latent Dirichlet Allocation to detect and interpret DNA-based biodiversity data sets; the data used in the study are from soil DNA samples. Roque et al. (2019) used the Latent Dirichlet Allocation to identify accident-related patterns by analyzing reports from Road Safety Inspections on major Irish roads.

In recent years, topic modeling has gained ground in business; for example, Saura et al. (2019) used topic modeling to identify key factors in content generated by Twitter users to create successful startups. Kanungsukkasem and Leelanupab (2019) present a model based on LDA, capable of combining textual data (especially news articles) and financial time series. Bastani et al. (2019) used consumer complaints from the CFPB (Consumer Financial Protection Bureau) to propose an intelligent system based on the LDA to automatically reveal consumer problems. Balakrishnan et al. (2020) used LDA to identify the main emerging topics based on textual analysis of digital payment applications and other machine learning techniques that investigated consumer sentiment. Zhang (2019) identifies 16 major consumer review topics on Airbnb using the LDA and assessed the impact of reviews on ads. Marcolin et al. (2019) used topic modeling to analyze hotel reviews in Porto Alegre city. The textual data set analyzed contained 23229 guest comments, collected on the TripAdvisor website, from 2011 to 2016.

In this dissertation, we consider the Latent Dirichlet Allocation (LDA) model as the basis of the study; the LDA is widely used in the data mining community for document classification. The LDA model is conditioned by three parameters: the Dirichlet hyperparameters (α and β) and the number of topics (K). Liu et al. (2020) state that despite the LDA being an unsupervised model and discovering topics automatically, the discovered topics do not always make sense to people, called the topic coherence problem. Cao et al. (2009) point out that choosing the best K identifies groups where the similarity is greatest within the cluster while the clusters are as small as possible. This enables a more explicit representation of the meaning of the topic. Arun et al. (2010) show that the challenge ensures that a small number of latent topics are sufficient to effectively represent a large corpus.

Good data analysis with topic modeling depends on a good selection of parameters to be effective, Liu et al. (2020) emphasizes two basic problems: from the viewpoint of the topic and from the viewpoint of the word. From the perspective of the topic, models can generate some meaningless topics and from the perspective of the word. However, we can understand the meaning of a topic by the words that are most likely to belong to the topic, confusing words such as 'fruit' may arise in basketball game topics.

Selection of the number of topics is important, especially in the business scenario. To illustrate the importance, figure 2 shows 10 customer reviews from department stores, extracted from the Better Business Bureau - BBB (*https* : *www.bbb.org*) website. Carrying out an analysis in customer reviews, it is possible to identify points regarding the exchange of products, prices, cards, cleaning and service. Analyzing 10 customer reviews is a super easy task due to data to be analyzed, but when the data goes beyond the hundreds, it is already a complicated task. Topic modeling comes as a tool to support this analysis and it is clear to support management decisions.

Customer 1	*Bought a defective tv and was refused a refund. Instead of apologizing I was told I should have paid extra their warranty.*
Customer 2	*Every time I pay my card they take like 2 months to report it to the credit bureau.*
Customer 3	*Great place and staff always try's to help us as best as possible. Good customer service.*
Customer 4	*I bought a tv on black friday and the tv is terrible. I wanted to return it and buy a better one but since I dont have the box they will not take it back.*
Customer 5	*I go into this store all of the time. It's always clean and I can find what I need.*
Customer 6	*Lower your prices. Treat all customers equally, with same pricing across all neighborhoods.*
Customer 7	*My card was charged for an item I purchased in June here it is now September.*
Customer 8	*Prices and promotions are good.*
Customer 9	*Store manager is the rudest person I've ever dealt with. Operator did not help. It's always hard to find someone when you need help.*
Customer 10	*The isles are usually clear with nothing on the floor and everything seems to usually be in order.*

Figure 2 – 10 customer reviews.
Source: Author.

In figure 3, we represent 3 choices of the parameter K in the LDA model for the data in figure 2. When a few topics are chosen ($K = 2$) this will produce results overly generic and

difficult to interpret. For example, Customer 1 is complaining about the exchange of products and Customer 5 is praising the cleanliness of the store, very different themes but grouped in the same topic.

When the opposite happens, i.e., the choice of many topics ($K = 15$), the result is few clusters and a dispersion of information. It is possible to observe that of the 15 topics, nine comprise a single customer review, i.e., there was no grouping and there is still Topic 5 (T5) that there are no classified customer reviews.

K = 2		T1	T2
Customer 1	Bought a defective tv ...	**0,515**	0,485
Customer 2	Every time I pay my ca ...	**0,508**	0,492
Customer 3	Great place and staff ...	**0,517**	0,483
Customer 4	I bought a tv on black ...	**0,508**	0,492
Customer 5	I go into this store all ...	**0,532**	0,468
Customer 6	Lower your prices. Tre ...	0,457	**0,543**
Customer 7	My card was charged ...	**0,521**	0,479
Customer 8	Prices and promotion ...	**0,560**	0,440
Customer 9	Store manager is the ...	**0,509**	0,491
Customer 10	The isles are usually c ...	0,464	**0,536**

K = 5		T1	T2	T3	T4	T5
Customer 1	Bought a defective tv ...	0,196	0,196	**0,232**	0,179	0,196
Customer 2	Every time I pay my ca ...	**0,224**	0,207	0,190	0,172	0,207
Customer 3	Great place and staff ...	0,210	**0,242**	0,177	0,177	0,194
Customer 4	I bought a tv on black ...	0,226	0,214	**0,238**	0,190	0,131
Customer 5	I go into this store all ...	0,182	0,167	0,212	0,197	**0,242**
Customer 6	Lower your prices. Tre ...	0,211	0,175	0,193	**0,228**	0,193
Customer 7	My card was charged ...	**0,225**	0,197	0,197	0,183	0,197
Customer 8	Prices and promotion ...	0,179	0,196	0,196	**0,250**	0,179
Customer 9	Store manager is the ...	0,171	**0,229**	0,186	0,200	0,214
Customer 10	The isles are usually c ...	0,200	0,169	0,185	0,215	**0,231**

K = 15		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
Customer 1	Bought a defective tv ...	0,066	0,111	0,066	0,051	0,051	0,066	**0,126**	0,081	0,051	0,066	0,051	0,051	0,051	0,051	0,066
Customer 2	Every time I pay my ca ...	0,067	0,051	0,067	0,082	0,082	0,082	0,067	0,067	0,051	0,051	0,051	0,067	**0,097**	0,051	0,067
Customer 3	Great place and staff ...	0,057	0,057	0,075	0,057	0,057	**0,092**	0,057	**0,092**	0,075	0,075	0,057	0,057	0,057	0,075	0,057
Customer 4	I bought a tv on black ...	0,060	**0,077**	**0,077**	0,060	0,060	0,060	0,060	**0,077**	0,060	**0,077**	0,060	0,060	0,060	**0,077**	**0,077**
Customer 5	I go into this store all ...	0,048	0,076	**0,105**	0,090	0,048	**0,105**	0,048	0,048	0,048	0,090	0,062	0,076	0,062	0,048	0,048
Customer 6	Lower your prices. Tre ...	0,070	0,086	0,054	0,070	0,070	0,054	0,054	0,054	0,086	0,054	0,054	0,054	0,070	0,070	0,102
Customer 7	My card was charged ...	0,061	0,075	0,061	**0,089**	0,075	0,061	0,061	0,047	0,047	**0,089**	0,047	0,075	0,075	0,061	0,075
Customer 8	Prices and promotion ...	0,063	0,052	0,052	0,087	0,075	0,052	0,040	0,063	**0,147**	0,063	0,123	0,052	0,040	0,052	0,040
Customer 9	Store manager is the ...	**0,076**	0,058	**0,076**	0,058	0,058	**0,076**	0,058	0,058	**0,076**	**0,076**	**0,076**	**0,076**	0,058	0,058	0,058
Customer 10	The isles are usually c ...	0,077	0,060	0,077	0,060	0,060	**0,095**	0,060	0,077	0,060	0,060	0,060	0,077	0,060	0,060	0,060

Figure 3 – Representation of different choices in the number of topics.
Source: Author.

When selecting $K = 5$ (figure 3) it is possible to clearly classify and identify the 5 topics. In Topic 1 (T1 - Customers 2 and 7) customer reviews related to cards are classified, probably stores have their own card. Topic 2 (T2) can be labeled as service, as customers 3 and 9 report iteration experiences with employees. Customers 1 and 4 report experiences related to exchanges of products purchased in stores, which leads to labeling Topic 3 (T3) as exchanges and returns.

In Topic 4 (T4 - Customers 6 and 8), customer reviews related to prices and promotions are classified. And in Topic 5 (T5) it can be labeled as cleaning and organization, as customers 5 and 10 report experiences with cleaning and organization of stores. Although the data is real, it is just an illustration to demonstrate the importance of correctly determining the number of topics, especially in business administration, as the information is valuable for business management.

Most topic modeling methods implicitly assume that the number of topics is known in advance, thus considering it an exogenous parameter. That is annoying, leaving the technique prone to subjectivities. The quality of insights offered by LDA is quite sensitive to the value of the parameter K, and perhaps an excess of subjectivity in its choice might influence the confidence managers put on the techniques results, thus undermining its usage by firms.

Lu et al. (2017) emphasizes that the choice of the number of topics is a sensitive parameter as it directly influences the ability to interpret the data. In the LDA model, it is a requirement that the number of topics be specified a priori. Although difficult, is important to choose the appropriate parameter values for the models.

Sbalchiero and Eder (2020) point out that the computational load increases proportionally to the number of documents, topics and terms in the corpus. This problem reflects a lot in the choice of the parameter; the metrics proposed to determine the number of topics can be time-consuming and intensive in terms of computation, which can be more than the execution of the LDA.

1.1 Research question

The LDA model has been successfully applied to a variety of study areas. Despite this, studies to determine the number of topics are still a little-explored field. Most topic modeling methods implicitly assume that the number of topics is known in advance. Determining the parameter K is extremely important and little explored in the literature, mainly due to the intensive and time-consuming computation procedure. In this sense, the research question is: Is there a faster way to subsidize objectively the decision on the number of topics to be considered in the LDA model?

1.2 Objectives

In this sense, this dissertation's main objective is to develop a metric to identify objectively the number of the parameter K of the LDA model that allows an adequate representation of the corpus and its implementation is an agile computing procedure.

To fulfill the main dissertation objective, we set these specific goals to achieve:

1. to survey the existing metrics to determine the parameter K in the LDA model;
2. to develop a novel technique to determine the parameter K;
3. to produce an algorithm to implement the novel method in a corpus;
4. to test the proposed technique efficiency and efficacy in two corpora;

5. to apply the existing metrics in two corpora;

6. to compare the proposed metrics with the proposed technique in those two corpora.

1.3 Dissertation organization

This dissertation is organized as follows. Section 2 is a literature review, organized in three parts: the first concerns a general approach to the Latent Dirichlet Allocation (LDA) model, the second evaluates the literature on LDA and the third reviews some related work in LDA and methods proposed to choose the number of topics. Section 3 is the main and innovative contribution of the dissertation. We present the proposal to determine the number of topics. Section 4 presents the experimental results of the proposed metric and other metrics. Finally, conclusions were drawn and a proposal for future work was presented in section 5.

2 THEORETICAL BACKGROUND

In this section, we describe the Latent Dirichlet Allocation (LDA) model proposed by Blei et al. (2003) is especially popular when discussing probabilistic topic models, given their importance and being the basis for developing other models found in the literature. The second part of this section provides a survey of work related to the LDA model. In the third part of this section, we deepen the studies in relation to the metrics for determining the K parameter.

2.1 Latent Dirichlet Allocation

A topic model is based on given a corpus of documents; it discovers the topics that permeate the corpus and assigns documents to these topics. Thus, one can think of a topic model as a box with two outputs: assigning words to topics and assigning topics to documents.

The first output, the topics, are distributions over words; as in figure 4, which shows three uncovered topics, topics are usually presented as word lists, that is, a handful of words. In practice, this is often enough to get a rough understanding of the topic.

In the second output of a topic, template documents are assigned to topics. In Figure 4 this step is represented by the simplex, where each of the seven documents is associated with the three topics, showing each document's position in the topic space.

There are various methods for finding these topics and assigning topics to documents; one of the main divisions of topic extraction models is non-probabilistic and probabilistic. Among the non-probabilistic topic extraction techniques, we have the Latent Semantic Indexing (LSI) proposed by Deerwester et al. (1990) where a matrix represents the word count per document and using singular value decomposition (SVD) to reduce the dimensions of the matrix preserving the similarity structure. Based on LSI, Hofmann (1999) proposes the probabilistic Latent Semantic Indexing (pLSI) and we now have the representation of the topic by a random variable sampled based on the words in the document. The pLSI is the first topic extraction technique using probabilities.

The pLSI model provides a good basis for researchers to perform a text analysis. However, Blei et al. (2003) highlights some problems. The first is in relation to the probabilistic model used only at the word level and not at the document level, i.e., the words are generated by random variables (multinomial distribution) and the documents are represented with a list of numbers and there is no generation of random variables for these numbers. Another problem pointed out by Blei et al. (2003) refers to the number of parameters with a positive linear relationship with the corpus, leading to problems of overfitting.

Aware of these problems, Blei et al. (2003) proposed the Latent Dirichlet Allocation

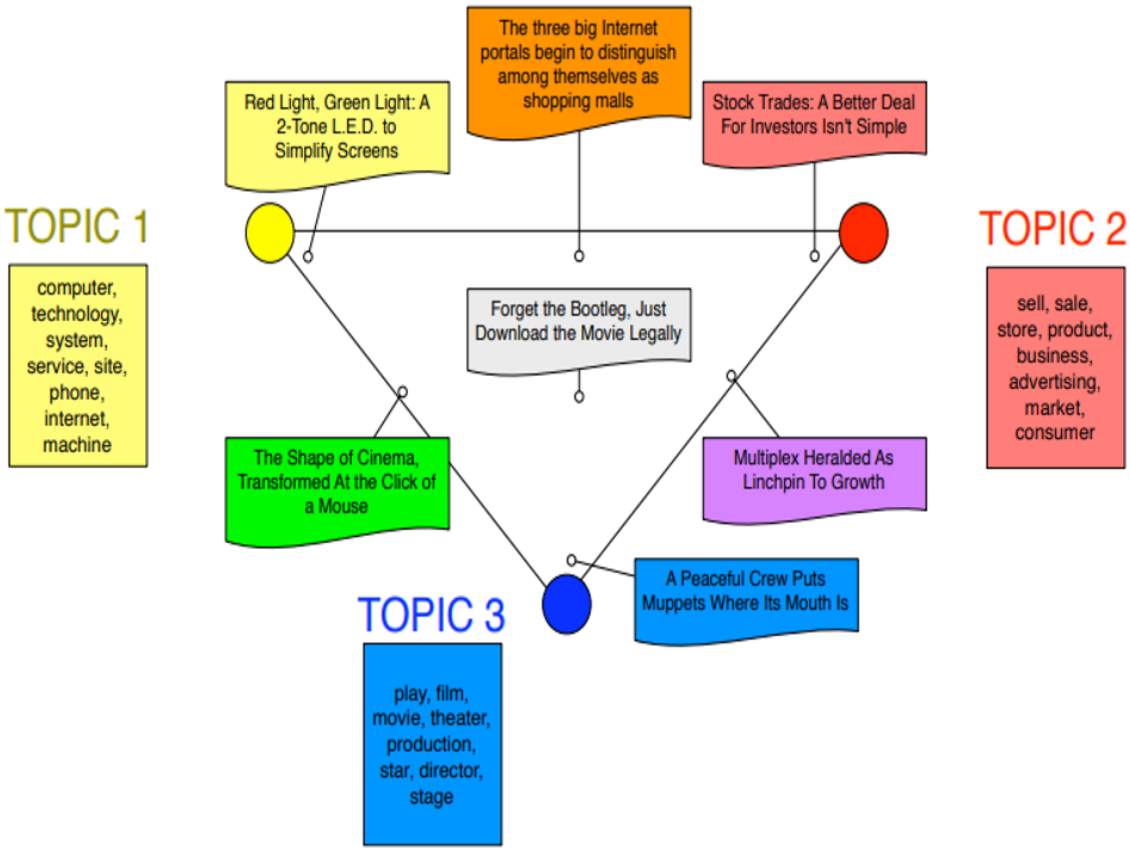


Figure 4 – Topics and Document Assignments to Topics.
Source: Adapted from Blei et al. (2003).

(LDA) model, which is the main, or the most popular, of the models of probabilistic topics.

Latent Dirichlet Allocation (LDA) is a document generating probabilistic model. In this model, the observable variables are the terms of each document, and the non-observable variables are the topic distributions. The parameters of the topic distributions, known as hyper-parameters, are given as priors in the model. The distribution used to sample the distribution of topics is the Dirichlet distribution. In the generative process, the Dirichlet sampling result is used to allocate the words of different topics, and they will fill in the documents. One can perceive the meaning of the name Latent Dirichlet Allocation, which expresses the intention of the model to allocate the latent topics distributed obeying the distribution of Dirichlet.

The intuitive idea behind the LDA and illustrated by the authors (BLEI et al., 2003) assumed that several topics, distributions over words, exist for the entire collection. Each document is assumed to be generated as follows: First, choose a distribution on the topics and then, for each word, choose a topic assignment and choose the corresponding topic word. In the seminal work, the authors exemplify the idea with a scientific paper on data analysis to determine the number of genes an organism needs to survive (in an evolutionary sense) by analyzing the article and different words used in the article as Computer, forecast, life, organism, genes, se-

quenced. In the example, if we take the time to highlight each word in the article, you would see this article combines genetics, data analysis, and evolutionary biology with different proportions. LDA is a statistical model of document collections that attempts to capture this intuition (BLEI, 2012).

The density function of the Dirichlet distribution, denoted as $Dir(z, \alpha)$, is presented in the equation 2.1:

$$Dir(z, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} {z_k}^{\alpha_k - 1} \tag{2.1}$$

Where $z = (z_1,...,z_K)$ is a K-dimensional variable, $0 < z_i < 1$ e $\sum_{i=1}^{K} z_i = 1$. Here, $\alpha_i > 0$ are the hyperparameters of the distribution. The $B(\alpha)$ function is the Beta function, which can be expressed in terms of the function Γ - equation 2.2:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{2.2}$$

Distributing Dirichlet has some important properties (BISHOP, 2006), and they are commonly used in Bayesian statistics; the distribution of Dirichlet is *a priori* conjugated of the multinomial distribution.

The LDA generator process is an imaginary process and, conversely to what is proposed in a computational task of extracting information, it is assumed the topics are specified before any data is generated. Here, topics are defined as probability distributions over a fixed vocabulary (words), while documents, nothing more than bags of words, arise from the probabilistic choice of words belonging to topic distributions.

The whole generative process can be represented graphically with a Bayesian network. This network is illustrated in figure 5.

As an introduction, the Bayesian model of LDA is a hierarchical model with three levels, where the first level represents the distribution of topics throughout the collection of documents. At the second level, one has the distribution of the topics for each document. And in the third level, distributing the topics internally for words in a document and with the last level is repeated, it is possible to represent a document as a mixture of topics.

To represent the distributions, two variables are used; the variable ϕ is an n-dimensional variable, where n is the number of words in the vocabulary. The variable θ is a K-dimensional variable, where the value of K is the number of topics. These two variables are generated by the Dirichlet distribution with their respective β and α hyperparameters.

Then, with the distributions ϕ and θ_d, the document d_j is generated. In the LDA model, itis considered that a document is simply a bag of words, with n_d terms in a document d. The terms of a bag of words are vocabulary words, and occasionally, repetitions of the same word

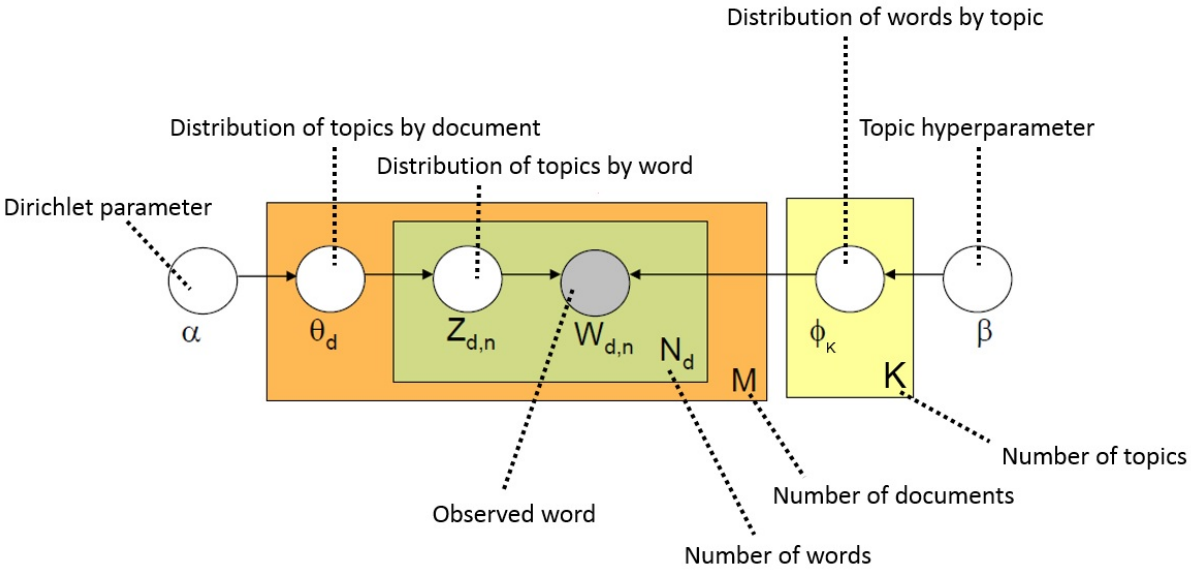


Figure 5 – Graphic model of the LDA.
Source: Adapted from Blei et al. (2003).

may occur. For each position i, of the n_d terms positions of a bag of words, a word of the distribution of topics is chosen. To do this, one must choose a topic k of the existing K topics and associate this topic with the position i of the document d.

The topic is chosen by obeying the θ_d distribution, which informs the participation of the topics in the d document; the variable z_{dn} will store the chosen topic. Then, the ϕ distribution is chosen as the word that will fill the position i. The variable ϕ are K n-dimensional distributions, where each distribution k, ϕ_k, corresponds to proportions of words that semantically describe the subject of which the topic k treats. The term w_{dn} should be chosen from the topic z_{dn}, obeying the word distribution $\phi_{z_{dn}}$.

In the LDA, the same document can be related to several topics with different proportions of relevance, since each document has its own distribution of topics θ_d. One can see this in the generative model by choosing the topic assigned to variable z_{dn}, where occasionally there will be a chance of choosing different topics according to θ_d distribution.

The generative process for LDA corresponds to the following joint distribution of the hidden and observed variables (2.3):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{dn}|\theta_d) p(w_{dn}|\beta_{1:K}, z_{dn}) \right) \tag{2.3}$$

Equation 2.3 determines a probability distribution with a complicated number of dependencies. The assignment of topic z_{dn} depends on distributing topics by document θ_d, and the observed word w_{dn} depends on the assignment of the topic z_{dn} and all topics $\beta_{1:K}$.

Considering the observed and unobserved variables, we aim to discover the assignments

of topics for the documents and the distributions of documents by topics and topics by terms. The big computational problem of LDA is to infer $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D})$, where w are all words observed in the document collection (BLEI et al., 2003). By the Bayes' theorem, we can formulate the probability as the calculation of the *a posteriori* of the LDA. We have the equation 2.4; the numerator is the joint distribution - equation 2.3 - of the model, and the denominator is the marginal probability of the observed data.

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{2.4}$$

The central computational problem can be solved by inferring the *a posteriori* probability of the whole model, described in equation 2.4. This can be thought of as the inverse of the generative process. Theoretically, this inference calculation can be done by summing the joint distribution of all possible values assigned to the unobserved variables (all words in the collection). The number of possible assignments is exponentially large, making this calculation computationally intractable (BLEI, 2012). Despite this, several methods approximate *a posterior* distribution and among the methods most used in the literature for inference of the LDA model is Gibbs Sampling (GRIFFITHS; STEYVERS, 2004).

The Gibbs Sampling is most popular mainly for the ease of implementation and its application in several problems. It is a special case of Monte Carlo simulation in Markov Chain. Markov Chain Monte Carlo methods can emulate high-dimensional probability distributions through the Markov chain stationary behavior (GEMAN; GEMAN, 1984).

The process performed by the Gibbs Sampling is based on sampling each dimension alternately, one at a time, conditioned to the value of all other dimensions. Suppose there is an unobserved K-dimensional variable, $z = \{z_1,z_2,...,z_K\}$, where z_i corresponds to the value of the ith vector dimension z and $z_{-i} = \{z_1,z_2,...,z_{i-1},z_{i+1},...,z_K\}$; also suppose the evidence given by the observed variable, w. In this example, the distribution to be inferred is $p(z|w)$. Instead of making samples throughout the z distribution to infer $p(z|w)$, the Gibbs Sampling makes separate choices for each i dimension of z, where sampling z_i depends on the other dimensions in z_{-i} sampled so far (RUSSELL; NORVIG, 2003).

With LDA, the Gibbs Sampling must sample three hidden variables, z , θ and ϕ. For simplicity, the Gibbs method applied in the LDA is collapsed to sample only the variable z, and from z find the values of the variables θ and ϕ. The sampling process is performed by statistics obtained by counting word assignments for topics and topics for documents made after sampling (BLEI, 2012).

2.2 Literature review on LDA

LDA is a generative statistical model, where based that each topic is a mixture of a set of words and that each document is a mixture of a set of topics. LDA is one of the most popular topic models for text analysis. This section provides a survey of the work related to the LDA model. First, analyzing the LDA in papers published in Brazil. Then analyzing the LDA in a broader collection of papers outside Brazil.

2.2.1 Brazilian Databases

For the Brazilian context, the bases were chosen: Scientific Electronic Library Online - SCIELO and SPELL Scientific Periodicals Electronic Library - SPELL. To complement the Brazilian research and justified by the area in which the dissertation is inserted, it was decided to consider the research in events carried out by *ANPAD - Associação Nacional de Pós-Graduação e Pesquisa em Administração*, *SOBRAPO - Sociedade Brasileira de Pesquisa Operacional* , *SBC - Sociedade Brasileira de Computação* and *ABEPRO - Associação Brasileira de Engenharia de Produção*. Finally, a search for theses was carried out at *BDTD - Biblioteca Digital Brasileira de Teses e Dissertações*.

The period considered is from 2003 to 2019 (2003 year of seminal work - LDA). First, the two terms "lda" and "latent dirichlet allocation" were used. After analyzing the results, some observed that the term "lda" is used for other areas such as chemistry, science, technology, mathematics and serves as an acronym for: Laser Doppler Anemometer (LDA), Left Displaced Abomasum (LDA), Limiting Dilution Assay (LDA), Linear Discriminant Analysis (LDA), Local Density Approximation (LDA), Loss Distribution Approach (LDA), Low Dose Aspirin (LDA), Lowly Disturbed Area (LDA) and Lung Densities Analysis (LDA). In this way, only the term "Latent dirichlet allocation" was a search term.

It can be seen from table 1 there are few studies on Latent Dirichlet Allocation published in Brazil.

Table 1 – National base.

Data base	Number of papers
SCIELO (http://www.scielo.br)	2
SPELL (http://www.spell.org.br)	0
SOBRAPO (http://www.sobrapo.org.br)	0
ABEPRO (http://portalabepro.educacao.ws)	0
SBC (http://portaldeconteudos.sbc.org.br)	0
ANPAD (http://www.anpad.org.br)	0
BDTD (http://bdtd.ibict.br/)	3
Total	5

The two articles (SCIELO base) found in the national context are Araújo (2017) and Moreira and Cesar (2019).

Araújo (2017), making use of network analysis and quantitative techniques of text analysis (Gibbs Sampling method, Bayesian algorithm derived from Latent Dirichlet allocation - LDA). The author used as corpus all Information Requirements (RIC) submitted by federal deputies of coalition parties between 1995 and 2014. The author has shown that in situations where ministerial portfolios are distributed to actors with different policy preferences, parties intensify the use of RICs to monitor the ministries and policies that interest them, besides listing the more controlled issues of government coalition parties. It also undermines the intra-cabinet control network and identifies that the parties responsible for portfolios with greater budget allocation are the actors with the highest degree of centrality in the networks of mutual monitoring.

The LDA model is used by Moreira and Cesar (2019) to analyze the materiality of the speech on a textual basis. They used the LDA to verify elements of speech, allowing quantifying the frequency and intensity. The basis for the study was the discourse of gender ideology, a discourse disseminated by conservative and fundamentalist groups. It was also possible to characterize the ESP Movement (from Escola Sem Partido, or School Without Party) as an impeller of the discourse of gender ideology on social networks.

The three dissertations found in the BDTD database are Santos (2015), Kaszubowski (2016) and Hardt (2019).

Santos (2015) proposes in its dissertation (Computer Science and Computacional Mathematics) the Latent Association Rule Cluster-based Model(LARCM), a non-probabilistic topic model that aims to represent the documents of a collection with reduced dimensionality, grouping the terms with information of the correlation between the terms. The LDA is used only as a means of comparison. However, as with other topic modeling methods, the K number of groups generated in LARCM must be reported by the user.

Kaszubowski (2016) proposed in his dissertation in psychology the creation of a model based on the LDA for the process of free association, a tool of psychoanalysis developed by Sigmund Freud in 1986, where the patient is guided to say what he is told come to mind. In the dissertation, the author successfully represents free associations with a topic model corroborated in the context of the inferences made for the case study. However, the author stresses that the volume of data used to adjust the model is not optimal.

Hardt (2019) used the LDA in his dissertation (International Relations) to analyze two corpus, the first with international news in Brazilian newspapers and the second composed of political speeches within Foreign Relations Commissions of the Brazilian Congress. The objective of the dissertation is to understand the dynamics between mass media and political speeches. The results show that the modeling of topics with LDA created coherent topics whose connection with real-world events can be verified. The author points out that results demonstrate that Brazilian politicians and newspapers are not isolated or unstable in relation to international issues.

Based on the results, it is observed based on the results that studies on the LDA model are scarce. In their entirety, they are studies with the applicability of the LDA model. And even considering the LDA model applications, they are still scarce with data from Brazil, further highlighting the importance of the study.

2.2.2 International Databases

For the international context, the bases were chosen: IEEE Xplore; Science Direct; Emerald; Springer; ACM Digital Library and Web of Science. The period considered is from 2003 to 2019. As evidenced in item 2.2.1 the term "lda" serves as an acronym in other areas, such as chemistry, science, technology, mathematics. In this way, only the term "Latent dirichlet allocation" was used as a search term.

As inclusion criteria, research articles published in newspapers or Proceedings of Conferences (this criterion is justified, as it is a characteristic of the computing area) and published in English were used. During the search, three articles were found that did not fit the above criteria, however they are recurrent citations in studies involving topic modeling (totaling more than 6 thousand citations according to Google Scholar metrics) and it was decided to include the articles.

Unlike the national context in the international context, it has a vast exploitation of the term "latent dirichlet allocation" (table 2).

Table 2 – International base.

Data base	Number of papers
Science Direct (http://www.sciencedirect.com/)	1723
Springer (https://www.springer.com/)	1462
Web of Science (http://apps.webofknowledge.com/)	1198
ACM Digital Library (http://dl.acm.org/)	463
IEEE (http://ieeexplore.ieee.org/)	165
Emerald (http://www.emeraldinsight.com/)	142
Others	3
Total	5156

During the research stage, we observed the existence of intersections between the bases, that is, the same papers was present in more than one database. Figure 6 summarizes these findings. After the analysis, 4495 papers were analyzed.

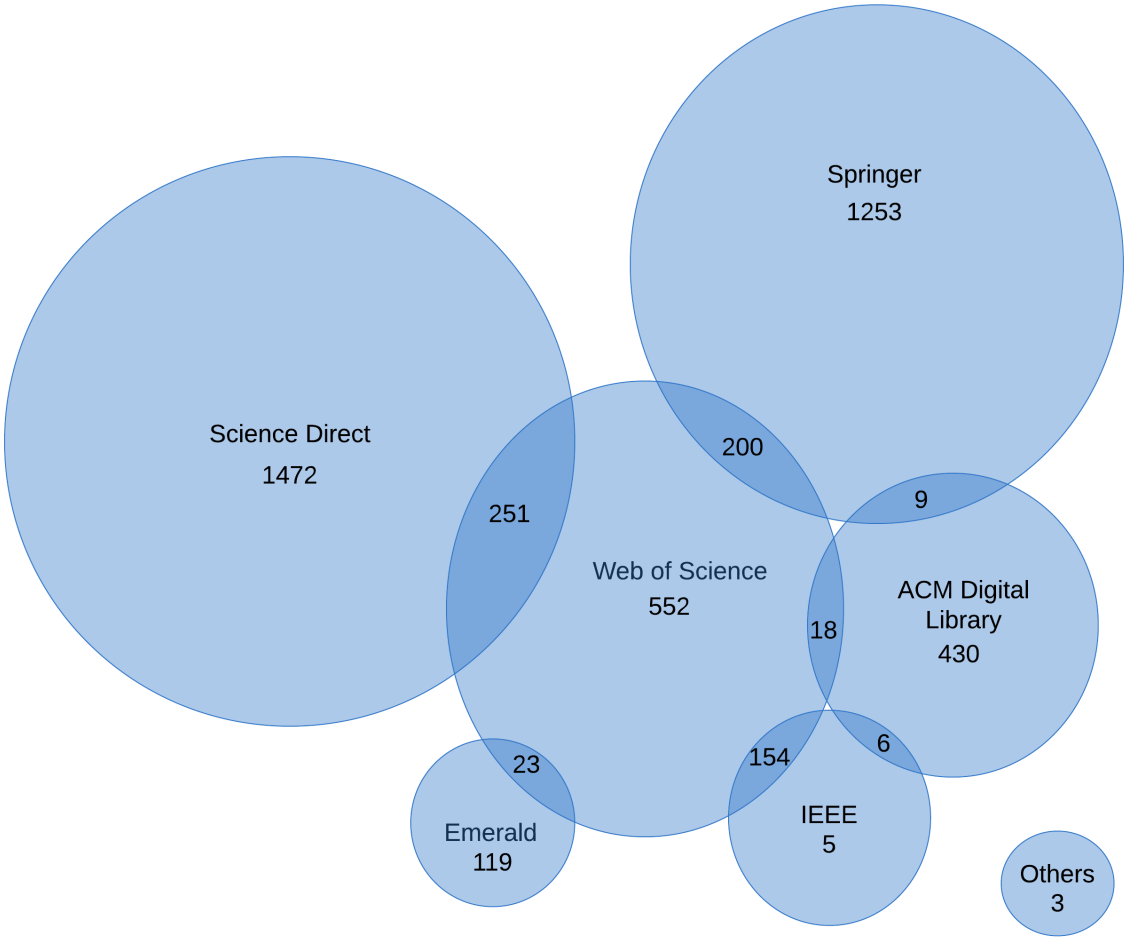


Figure 6 – Intersections between databases.
Source: Author.

The number of papers published on the subject per year (figure 7) started shyly being counted in dozens until 2011 and follows an increasing trend. In 2019 the concentration of papers is 23% (1034 papers) of the total found in the search (4495 papers).

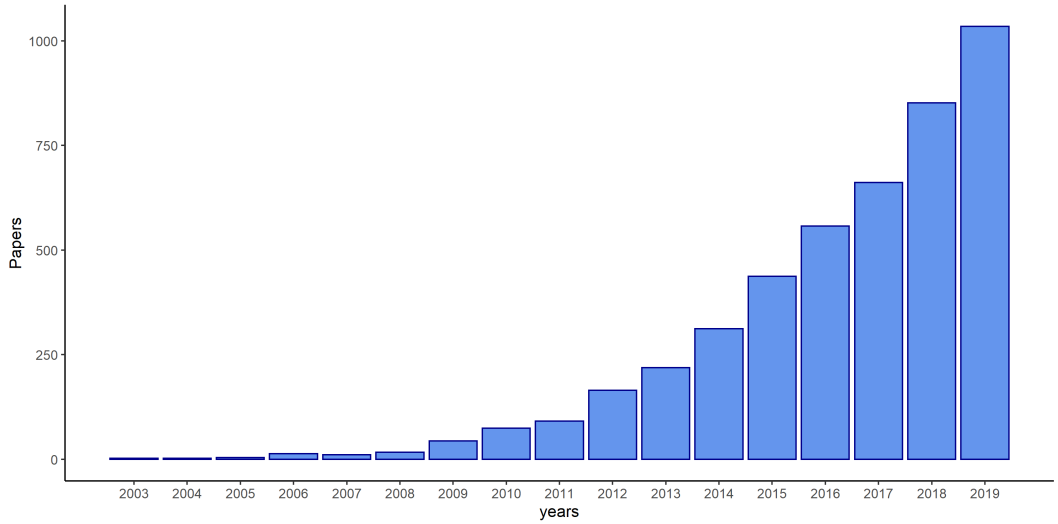


Figure 7 – Annual evolution of publications related to LDA (2003 - 2019).
Source: Author.

In table 3, which lists the 10 journals that most published the papers, it was noticed that the Latent Dirichlet Allocation technique is applied in several journals because these 10 journals present contain only 21.85% of the total found. Emphasizing the diversification of the area where the technique is applied.

Table 3 – Distribution of papers by journals.

Journals	Number of papers	%
Expert Systems with Applications (ISSN: 0957-4174)	130	2.89
Neurocomputing (ISSN: 0925-2312)	129	2.87
Multimedia Tools and Applications (ISSN:1573-7721)	115	2.56
ACM Transactions on Management Information Systems (ISSN: 2158-656X)	106	2.36
Knowledge-Based Systems (ISSN:0950-7051)	104	2.31
Procedia Computer Science (ISSN: 1877-0509)	91	2.02
Information Processing & Management (ISSN: 0306-4573)	89	1.98
Scientometrics (ISSN: 0138-9130)	87	1.94
Information Sciences (ISSN: 0020-0255)	76	1.69
Empirical Software Engineering (ISSN:1382-3256)	55	1.22
Others	3513	78.15
Total	4495	100.00

After the surveying the bases, all paper abstracts were read and classified into 12 categories (categories were defined for the researcher's convenience): literature review; determination of the number of topics; language/semantic; big data; system of recommendations; images; automation system; social networks; discovery of knowledge; topics in time; classifications/-groupings/applications; and analysis of feeling. Due to the large number of papers, it was decided to comment on 5 papers of each category, which would be the articles that most represent the category in which they were classified (table 4). This classification and the choice of arti-

cles to be commented were at the discretion of the researcher, considering the interest of the proposed research.

Table 4 – Division of papers by categories.

Category	Number of papers	%
1. System of recommendations	742	16.51
2. Classifications/groupings/applications	652	14.51
3. Language/semantic	588	13.08
4. Images/videos	577	12.84
5. Big data	495	11.01
6. Discovery of knowledge	340	7.56
7. Topics in time	237	5.27
8. Sentiment analysis	234	5.21
9. Automation system	234	5.21
10. Literature review	207	4.61
11. Social networks	182	4.05
12. Determination of the number of topics	7	0.14
Total	4495	100.00

Based on the table 4 and the analysis of articles in their classifications (which will follow), the large and diversified applicability in the LDA model is evident. It has been applied to a variety of applications and also serves as building blocks in other powerful models. Despite this, studies aiming to determine the number of topics is still a little-explored field.

1. System of recommendations - In this set, we selected papers that deal with recommendations systems, group of techniques and algorithms that selects items based on interaction data and user interests. These recommended items can be of various types like books, movies, music, videos, products in an e-commerce store. In this sense, topic modeling enters Group Analysis with grouping a data set so the objects belonging to the same group are more similar (according to some criteria) between them than those in other groups. Kim and Saddik (2015) addresses the issue of recommending items for communities of interest (i.e., groups) that are specifically formed in social media systems. Lai and Hong (2017) also addresses the issue recommendation for groups. Bellogín et al. (2013) we present a comparative study about the influence that different types of information available in social systems have on the recommendation of items. Other work in the direction of recommendation: He et al. (2016); Zhao et al. (2018); Mukherjee et al. (2015); Ma et al. (2017); Saraswat et al. (2016) and Zheng et al. (2014).

2. Classifications/groupings/applications - In this set were selected papers that deal with classifications, groupings and applications. Layman et al. (2016) applied LDA to a corpus of NASA problem reports to extract trends in tests and operational failures. Topical

modeling can identify problematic issues within missions and during mission lives, providing useful feedback to engineers and project managers. Christidis et al. (2012) have used LDA to get hidden topics and use the latter to evaluate similarities in recommending features and tags, as well as expanding query results. Ye et al. (2016) extracted and constructed a list of dengue-related keywords to analyze how often these words appear in Weibo messages based on the LDA. Subsequently, they applied spatial analysis to detect how cases of dengue cluster spatially and spread over time. Kim et al. (2011) show how to discover clusters of abuse tasks using LDA. They have applied LDA to hundreds of thousands of untitled job postings from Freelancer.com, identified that he discovers clusters of related abuse work and identifies the dominant words that distinguish them. Using LDA clusters to create a profile of the population of workers who bid on abuse jobs and the population of buyers who post descriptions of their projects. Other work towards groupings and applications: Wenming et al. (2016); Perina et al. (2010) and McLaurin et al. (2018).

3. Language/semantic - In this set were selected papers that deal with language and semantic. Bost et al. (2015) present an automatic analysis of real-life phone conversations between an agent and a customer in a customer service department. Chen et al. (2018) present an automatic construction method for the advertising ontology. The construction method searches for related Web documents, extracts keywords, and ponders keywords for concepts. Other works can be cited as Brychcín and Konopík (2014); Sta et al. (2016); Cahyaningtyas et al. (2017) and Zhang et al. (2018).

4. Images/videos - In this set were selected papers that deal with the classification of images. Shang and Chan (2011) in this paper the authors extended the LDA model to model the dynamics of facial expression. The proposed model integrates the temporal information of sequences of images through the redefinition of the probability of generation of topic without involving new latent variables or to increase the difficulties of inference. Nie et al. (2017) applied the LDA model for the extraction of visual topics and we used the visual image distribution feature to deal with the problem of 3D object retrieval. The LDA model is used to extract the topic template to deal with the recovery problem. Other works using LDA or LDA adaptations: Zeng et al. (2015); Zhang et al. (2017) and Arun and Govindan (2018).

5. Big data - In this set were selected papers that deal studies with big data. Bolelli et al. (2009) propose a generative model based on LDA for mining of distinct topics in collections of documents, integrating the temporal ordering of documents in the generative process. The experiments were carried out in the collection of scholarly works of the CiteSeer repository. Huang et al. (2014) explored a hybrid intrusion detection approach by discovering knowledge from big data using LDA. They pointed to the "hidden" patterns of operations performed by normal users and malicious users of a large volume of network

logs. Sukhija et al. (2016) explore the use of digitized social science data, text analysis tools to generate topic models, view topics to strengthen intersectional research engaging the relationship between consumption, race, class, and gender in sociology. Other works can be cited as Blazquez and Domenech (2018) and Sekiya et al. (2010).

6. Discovery of knowledge - In this set were selected papers that deal with the discovery of knowledge - is a process that allows extracting knowledge of information stored in large specialized databases. Yeganova et al. (2018) present a contribution at the theoretical and algorithmic levels, as well as demonstrate the viability of the method for large-scale applications. Bimba et al. (2016) presents the survey of publications related to knowledge base modeling and manipulation technologies, between the years 2000-2015. Other works can be cited as Papachristopoulos et al. (2015); Liang et al. (2018) and Huang et al. (2014).

7. Topics in time - He et al. (2019) created a system divided into modules with different methodologies highlighting ARIMA to analyze the public opinion index and opinion analysis based on an LDA thematic model and a word cloud map. The main objective of the study was to improve the performance of the analysis of opinions and comments on public online bases in colleges and universities using short-term trend forecast results. Ma et al. (2019) applied a system to the LDA to decrease the perplexity of the models and increase the quality of the clusters, the system is based on time series, improving the discovery of topics in short texts. Chen (2017) proposes a method based on the LDA for research on blogs considering the time factor characteristics such as the popularity of the post and time of the post are considered. Zhang et al. (2018) used the modified LDA model to identify users preferences by inserting the time factor with weight assignment to each classification according to the date and time. Wilson et al. (2018) proposes a model based on LDA to represent the temporal behavior of the profile of communication service customers.

8. Sentiment analysis - seeks to identify the feeling that users present about some entity of interest (a specific product, a company, a place, a person, among others). Fujimoto et al. (2011) proposes a structure of profile creation and clustering of Web users based on LDA-based topic modeling, with an analogy to document analysis, in which documents and words represent users and their actions. Yang et al. (2015) focus on contextual suggestions based on location and propose to leverage user opinions to build profiles. Instead of logging "what a user likes or does not like" he proposes a template to identify "why a user likes it or not" to better predict whether a user would like a new one. Srinivasan et al. (2017) seek to understand the "interest" of Twitter users, where we define interest as the maximum number of tweets a user posted on a single topic. Other works may be cited: Hu et al. (2017); Chen and Ren (2017) and Al-Obeidat et al. (2018).

9. Automation system - In this set were selected papers concerned with the automation of

topic processing. Song et al. (2016) propose a context recognition mechanism to acquire Experiential knowledge (EK) automatically and in a timely manner. The proposal comprises a formal description of EK using standard ontology and logic, a machine learning method that uncovers questions and answers from the context of collaborative engineering tasks, and a semantic mapping step transforming the discovered questions and answers into concepts and ontological relationships. Cerisara (2009) deals with automatic lexical acquisition and topic discovery from a speech stream. The proposed algorithm constructs a lexicon enriched with topic information in three steps. Other works can be cited as Asnani and Pawar (2017); Troussas et al. (2017); Wang et al. (2012) and Escalante et al. (2012).

10. Literature review - In this set were selected papers concerned with literature review referring to several areas. Bao and Datta (2014) implemented a variation of the LDA model to simultaneously discover and quantify risk types from textual disclosures of risk. Experimental results show that our proposed method overcomes all competing methods and can find more significant topics (types of risk). Grimaldi et al. (2017) present a systematic review of the literature and a synthesis of high-quality contributions focusing on an overview of Open Innovation (OI) research and Intangible Assets (IAs). Mäntylä et al. (2018) present a computer-assisted bibliographic review, where use text mining and qualitative coding and analyze 6996 Scopus articles. The authors point out that the roots of the analysis of feeling are in the studies on analysis of public opinion in the early twentieth century and in the text analysis of subjectivity performed by the community of computational linguistics in the 1990s. Other works can be cited as Nazar et al. (2016); Kappassov et al. (2015) and Nassirtoussi et al. (2014).

11. Social networks - In this set were selected papers concerned with focused on studying the social networks. Xie et al. (2015) propose a forensic analysis method for an e-mail network based on the LDA topic model and the Centralization algorithm. The method considers the content of the e-mail and the communication structure by e-mail. Alp and Öğüdücü (2015) develop a method to overcome the challenge of tweets being too short for topic modeling. Compared different outline modeling schemes based on LDA. Sun and Lee (2017) propose a framework for recommending k-tours to meet user interest and time using user-generated content on a social network of photo sharing. They used the LDA model to categorize the hashtags posted by users on reference topics and then use those topics to characterize landmarks and users. Other works can be cited as Whelan et al. (2016); Wang et al. (2014) and Steiger et al. (2015).

12. Determination of the number of topics - In the studies presented in the other 11 categories, determining the number is often obscure in the papers. Some authors suggest that the value is known *a priori*, by model training system or simply by convenience. The studies found related to determining the K parameter are Griffiths and Steyvers (2004); Cao et al.

(2009); Arun et al. (2010); Deveaud et al. (2014); Cheng et al. (2015) Zhao et al. (2015) and Albuquerque et al. (2019). The studies are described in detail in the item 2.3.

2.2.3 Literature review on LDA in business

The discovery of knowledge from the text content has been a commodity much sought after by companies (from different areas) and has narrowed the relationship between companies and academic research.

In marketing, one of the areas that most uses topic modeling, there are several applicabilities such as: Pournarakis et al. (2017) presents a computational model that extracts subjects from consumers perceptions on social networks. Producing insights for brand recognition and brand meaning.

There are several applications of topic models in customer satisfaction surveys: for example, Gao et al. (2016) used topic modeling to analyze 17,747 comments from public transport customers in the United States. According to them, the most frequently identified customer satisfaction attributes are: waiting time, cleanliness, accessibility, comfort. Lucini et al. (2020) analyzed 55,775 Online Customer Reviews (ORCs) from passengers from different countries and airlines and identified 27 dimensions of customer satisfaction. Liao et al. (2020)focused on insurance companies in the USA and they used topic modeling to analyze information from customer calls. They classify and process the information more efficiently by using approximately 10,000 customer calls to develop the study. Sutherland et al. (2020) used 104,161 online reviews from Korean accommodation customers to identify which topics are considered important by guests. The researchers identified 14 topics such as accessibility, hospitality, room size, among other things.

In Managing Human Resources, Jung and Suh (2019) used LDA to identify and analyze factors for job satisfaction from employee reviews and identified that senior management is the most important factor for overall job satisfaction and provided insight into the decision making of decisions in managing employee satisfaction. Madding et al. (2020) worked with LDA to develop an IT professional profile that can help create an advantage for recruiters. Using linkedin data, the survey provides information that employers can use to address and meet the demands of an ever-changing job market, such as technology. Jayaratne and Jayatilleke (2020) used data from an online chat-based interview tool. The data are from an interview questionnaire with questions about previous experiences, situational judgment and values of the candidates. The researchers present a study based on the LDA model to classify and infer someone's personality from the use of language in a recruitment interview environment.

In finance, Calomiris and Mamaysky (2019) studied the classification and content of news articles to predict stock market risk and return in 51 countries. Using topic modeling, they classified and identified that the news divides into good or bad news relevant to both returns

and risk. BROWN et al. (2020) used LDA to identify and quantify what is being disclosed in the annual financial statement reports as opposed to how it is being disclosed, to detect incorrect financial reports. Mokhtari et al. (2020) to improve the prediction of financial reporting comment classes used the LDA model.

Schmiedel et al. (2019) studied Topic Modeling as a Strategy of Inquiry in Organizational Research and emphasizes that topic modeling cannot extract information about text data in an automated way at the touch of a button. Researchers need to make many decisions throughout all stages of their study and require more subjective interpretations.

Research related to Operational Research (OR) is still timid and there is scope for research connecting OR, business and data mining (MORTENSON et al., 2015). Baechle et al. (2020) are concerned with predictive models that add value to hospitals, reducing costs and penalties associated with preventable patient readmissions. Using text data from 16 hospitals as the primary data source and applying topic modeling, they created a model to optimize the cost of readmission.

Hindle et al. (2020) emphasizes that with the growth of business analysis and data science in recent years the role of OR is being increasingly challenged. Hindle et al. (2020) states that with the rapidly evolving world of technologies, the field of business analysis is gaining maturity and consolidating itself as an applied science in real organizational problems requiring an ability to structure problems, a space in which operational research (OR) must find a natural home.

2.3 Number of topics and topic selection

The LDA model is conditioned by three parameters: the Dirichlet hyperparameters (α and β) and the number of topics (K). The choice of parameter K has important implications for the results produced by the model, since a relatively large number assigned to parameter K can disperse text allocation, and a relatively small number assigned to parameter K can condense text allocation too much, hindering a clear analysis.

The choice of the appropriate value for K is a model selection problem, addressed through a standard Bayesian statistical method. The answer is to calculate the subsequent probability of this set of models, given the observed data. Requires the training of several LDA models (one for each value of K) to select one with the best performance. It is an intensive and time-consuming computation procedure.

In the seminal paper Blei et al. (2003) they use *perplexity* to compare the performance of pLSI with LDA. Subsequently, *perplexity* was used by some authors to determine the value of K.

For a data set D_{test}, the *perplexity* is that defined in the equation 2.5:

$$perplexity(D_{test}) = exp\left\{-\frac{\Sigma_d log(P(w_d))}{\sum_d N_d}\right\} \tag{2.5}$$

where w_d is a word and N_d is the number of words, both are from the test dataset.

When calculating the *perplexity*, a part of the data is reserved for testing purposes and the rest used for training the models (quantities determined at the researcher's convenience). The goal is to find a low *perplexity* score, as it would indicate a model with generalization performance (KIM et al., 2011).

Wenming et al. (2016) explains that intuitively the *perplexity* will be lower when the model is more likely to have words in the document than words that are not. Kim et al. (2011) further states that the reliability of the result obtained in *perplexity* will depend on the amount of data training. Chang et al. (2009) points out that *perplexity* will indicate a topic value consistent with human evaluation. The *perplexity* will not be considered in the study for comparison purposes because it requires data training.

In Griffiths and Steyvers (2004) the proposal is to produce a set of models by changing the K parameter. The focus is to analyze P($w|K$), where w is the word. Griffiths and Steyvers (2004) points out that the complication is to obtain the sum of all possible word assignments to topics z. They solve using an approximation of P($w|K$) obtaining the harmonic mean of a set of these values of P($w|z$,K) through the equation 2.6.

$$P(w|z) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + W\beta)} \tag{2.6}$$

where $n_j^{(w)}$ is the number of times the word w has been assigned to topic j in the assignment vector z and $\Gamma(\cdot)$ is the standard gamma function.

Griffiths and Steyvers (2004) describes the behavior of the metric that increases until it reaches its peak and then decreases, a profile often seen when varying the dimensionality of a statistical model. The big problem is precisely the processing time of the models to later calculate the suggested metric.

Cao et al. (2009) proposes a metric that considers the correlation between topics. The average cosine distance between each pair of topics is used to measure the stability of the topic structure.

$$ave\_dis = \frac{\sum_{i=0}^{K} \sum_{j=i+1}^{K} corre(T_i, T_j)}{K \times (K-1)/2} \tag{2.7}$$

Given a topic Z and the distance r, calculating the average cosine distance between Z

and the other topics, the number of topics in the radius of r of Z is the density of Z, called $Density(\mathrm{Z,r})$.

The average cosine distance of the model $r1$ = ave_dis (β), the densities of all topics Density (Z, $r1$) and the cardinality of the old model $C = Cardinality(LDA, 0)$ are calculated sequentially. The model is re-estimated based on Cardinality $K_{n+1} = K_n + f(r) \times (K_n - C_n)$. r is guided by $f(r)$. If it is negative $f_{n+1}(r) = -1 \times f_n(r)$, else $f_{n+1}(r) = -f_n(r) \cdot f_0(r) = -1$.

After calculating the density of each topic, we find the most unstable topics in the old structure and iteratively update the K parameter until the model is stable. The processes are repeated until the average cosine distance and cardinality of the LDA model converge. Searching for the minimum value, indicating the best K.

Arun et al. (2010) proposes to consider the information of the word-topic and also document-topic, unlike Cao et al. (2009) which considers only word-topic. Two matrices: M1 (of the topic and document order) and M2 (of the document and topic order) result from the matrix factoring of a document and word frequency matrix. The proposed metric calculates the Kullback-Leibler symmetric divergence of the Singular value distributions of the M1 and M2 matrix.

$$ProposedMeasure(M1, M2) = KL(C_{M1} \parallel C_{M2}) + KL(C_{M2} \parallel C_{M1}) \tag{2.8}$$

Where, C_{M1} is the distribution of singular values of the M1 matrix (Topic-Word). C_{M2} is the distribution obtained by normalizing the vector L $\times$ M2 (where L is 1 $\times$ D of the length of each document in the corpus and matrix M2 (Document-Topic).

The objective of the metric is to find the lowest value, because the higher the number of the divergence, the lower probability values for words that do not belong to a topic occur.

Deveaud et al. (2014) proposes a method for mining and modeling latent research concepts called Latent Concept Modeling (LCM), but the method depends on using the LDA to display highly specific topics related to user research. They look for the estimated number of topics together with their associated topic model, the idea is to find the model where the number of topics is more dispersed. The number of concepts is given by the equation 2.9.

$$\hat{K} = \underset{K}{argmax} \frac{1}{K(K-1)} \sum_{(k,k') \in \mathbb{T}_K} D(k \parallel k') \tag{2.9}$$

$\mathbb{T}_K$ is the set of modeled K topics. Deveaud et al. (2014) uses the Jensen-Shannon diver-

gence to calculate divergences between all pairs of topics.

$$D(k \parallel k^{'}) = \frac{1}{2} \sum_{w \in \mathbb{W}_k \cap \mathbb{W}_{k'}} P_{TM}(w \mid k) log \frac{P_{TM}(w \mid k)}{P_{TM}(w \mid k^{'})} + \frac{1}{2} \sum_{w \in \mathbb{W}_k \cap \mathbb{W}_{k'}} P_{TM}(w \mid k^{'}) log \frac{P_{TM}(w \mid k^{'})}{P_{TM}(w \mid k)} \tag{2.10}$$

$\mathbb{W}_K$ is the set of n words with the highest probabilities $P_{TM}(w|k) = \phi_{k,w}$ in the topic K:

$$\mathbb{W}_k = \underset{w}{argmax}[n]\phi_{k,w} \tag{2.11}$$

The metric proposed by Deveaud et al. (2014) always seeks the highest values.

The metric proposed by Cheng et al. (2015) provides a lower limit and an upper limit for the number of topics K in the LDA model. The upper limit is obtained by limiting the difference between M_2 (equation 2.12) and $\widehat{M}_2$ (equation 2.13). When the upper and lower limits converge is where there is a reduction in the range of possible K.

$$M_2 = \sum_{k=1}^{K} \frac{\alpha_k}{(\alpha_0 + 1)\alpha_0)} \mu_k \otimes \mu_k \tag{2.12}$$

$$\widehat{M}_2 = \frac{\sum_d \sum_{\ell \neq \ell'} X_{d\ell} \otimes X_{d\ell'}}{DL(L-1)} - \frac{\alpha_0}{\alpha_0 + 1} \widehat{M}_1 \otimes \widehat{M}_1 \tag{2.13}$$

Where d is the number of documents, ℓ number of documents, $X_{d\ell}$ is the word number in the document, K is the number of topics, $\alpha_0 = \sum_{k=1}^{K} \alpha_K$ and $\widehat{M}_1$ is described in the equation 2.14 (CHENG et al., 2015).

$$\widehat{M}_1 = \frac{\sum_d \sum_\ell X_{d\ell}}{DL} \tag{2.14}$$

Zhao et al. (2015) uses a heuristic approach to determine the ideal number of topics. The focus is to find analyzing the so-called perplexity change rate (RPC), where the RPC-based change point is determined as the most appropriate number of topics. Candidate numbers for topics are ranked in ascending order. For each number of candidate topics an LDA model is built m times in a training set combining $m-1$ subsets of the entire data set. For each candidate number of topics, the average perplexity of the test set number is considered. The rate of change of perplexity RPC is calculated using the equation 2.15:

$$PRC(i) = \left| \frac{P_i - P_{i-1}}{t_i - t_{i-1}} \right| \tag{2.15}$$

Where P are the average perplexities and t is the number of topics. The ideal number of topics corresponds to the first i that satisfied $RPC(i) < RPC(i+1)$, that is, analyzing graphically is the point where there is a change in the inclination of RPC versus number of topics (ZHAO et al., 2015).

The metric proposed by Albuquerque et al. (2019) is also based on the graphical analysis. The result is found by analyzing the graph of a matrix theta, where the parameter K is chosen where there is a drop in the probability associated with each cluster. The theta matrix is a $C \times S$ matrix, where S comes from a Y matrix with dimension equal to $L \times S$ where each row represents a sampling unit and each column a variable that describes these elements. O comes from a latent matrix Z with dimension equals to $L \times C$ where each row represents a sampling unit L and each column a possible state or cluster C. The matrices are built on top of a sampling of the original data, the size of this sampling is defined through the heuristic research principle Occam's razor where the objective is to create the smallest possible number of clusters, which is achieved by assuming a truncated stick-breaking prior, equation 2.16:

$$\theta_{lc} = V_{lc} \prod_{c^*=1}^{c-1} (1 - V_{lc^*}) \tag{2.16}$$

where $V_{lc} \sim Beta(1, \gamma)$ for $c = 1, ..., C-1$ and $V_{lC} = 1$ by definition.

Finally, the table 5 presents a summary of the proposals for obtaining the number of topics. The metrics were classified according to the decision base and the metrics where the decision base is punctual and the decision point (maximum or minimum).

Table 5 – Summary of techniques for determining the parameter K.

Papers	Decision base	Decison point
Griffiths and Steyvers (2004)	point	maximum
Cao et al. (2009)	point	minimum
Arun et al. (2010)	point	minimum
Deveaud et al. (2014)	point	maximum
Cheng et al. (2015)	graphical	-
Zhao et al. (2015)	graphical	-
Albuquerque, Valle and Li (2019)	graphical	-

Each method has its characteristics to indicate the value of the parameter K, the problem of the metrics found in the literature is the processing time, as will be evidenced in the next sections.

3 THE PROPOSED METHOD FOR LDA MODEL SELECTION

Based on the references, we observed that there are several metrics to determine the best K parameter in the LDA model providing different results. These efforts aim to provide a less subjective estimation of a value for the K parameter.

One of the most questioned points after applying the proposed metrics is their processing time, an important issue for algorithm efficiency. We propose a new method to estimate the best K parameter which produces results as good as the current methods much more quickly, therefore being more efficient computationally.

The proposed method is described as follows:

(1) Run LDA with a large K value. This is called our base model. The base model produces a set of K topics,$\mathbb{T}$.

(2) For the base model, estimate the probability of each word in each topic, $P(w|k)$, $k = 1,2,,K$. We have that $\sum_{k=1}^{K} P(w|k) = 1$ for each w in the corpus.

(3) The topics on the base model are then grouped into an ordered sequence of topics, $T_n = \{k \in \mathbb{T} | k \leq n\}$

(4) For each n, we compute

$P(w|T_n) = \sum_{k=1}^{n} P(w|k)$ and

$PL(T_n) = ln\left(\frac{nw}{\sum_{w=1}^{nw} \frac{1}{P(w|T_n)}}\right)$

where nw is the number of words in the corpus.$P(wT_n)$ is the probability that the word w belongs to group T_n and $PL(T_n)$ is the harmonic mean of $P(w|T_n)$.

(5) Finally, we seek the value of n that maximizes $PL(T_n)$. That is n^* is such that $PL(T_{n^*}) = max_n PL(T_n)$

We therefore have: Best $K = n^*$.

In figure 8 the proposed method is illustrated intuitively. The topics (columns) and words (lines), the proposed method groups the topics of the subsequent topic to carry information ($P(w|k)$) from the previous topics, resulting in the $P(w|T)$.

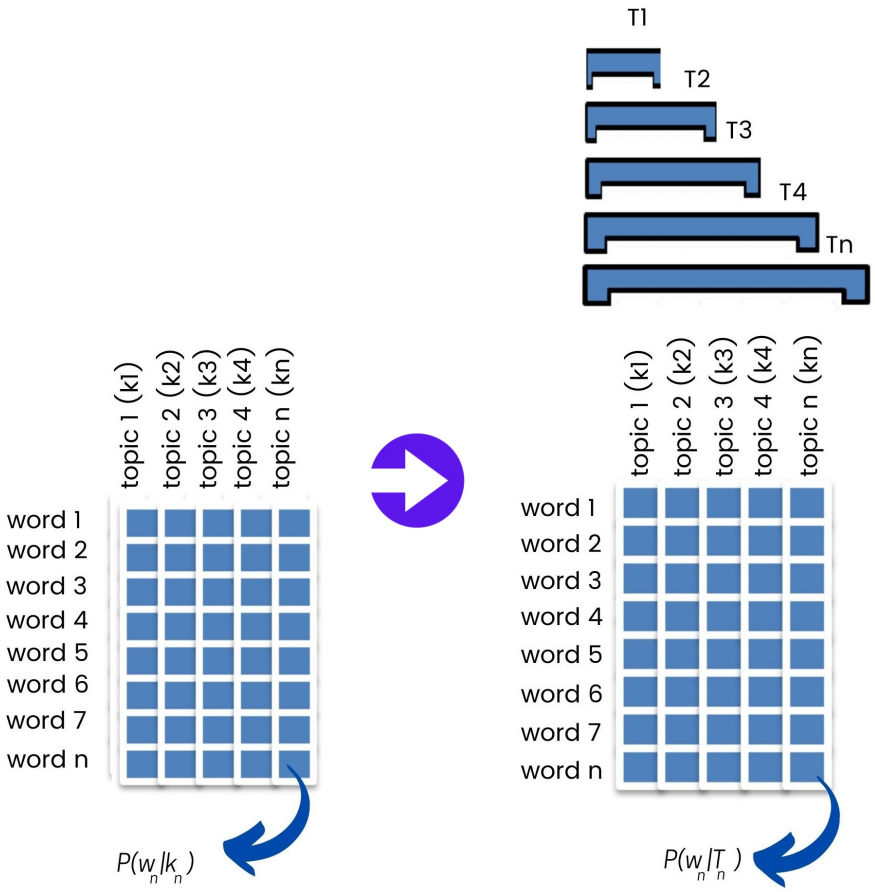


Figure 8 – Intuitive representation of the proposed method.
Source: Author.

This grouping facilitates and reduces the processing time, since intuitively the information for each topic (K) is loaded in the topic T. Subsequently, the harmonic mean of $P(w|T)$ is extracted and the point of greatest of harmonic mean value is located in the set. Because the property of always being the smallest of the three Pythagorean means(except in the limit case $a = b$, when the three averages coincide), the harmonic mean was chosen. Other properties that were considered when choosing the harmonic mean: the fluctuations of the observations do not affect the harmonic maan e more weight is given to smaller items (BOYER; MERZBACH, 2011; AGARWAL, 2020).

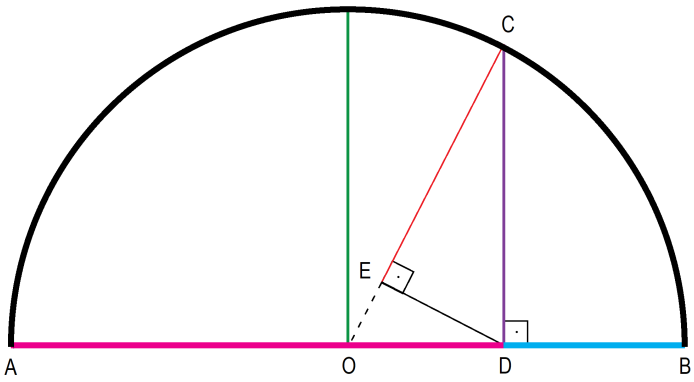


Figure 9 – A geometric construction of the three classical Pythagorean means.
Source: Adapted from Agarwal (2020).

In figure 9 is the geometric representation of the arithmetic mean, geometric mean and harmonic mean, where: $\overline{AD} = a$, $\overline{DB} = b$, $\overline{OC}$ = arithmetic mean of a and b, $\overline{CD}$ = geometric mean of a and b and $\overline{CE}$ = harmonic mean of a and b (AGARWAL, 2020).

The proposed method is defined as pseudo-code in Algorithm 1 and the code implemented in R is described in in Appendix A.

Algorithm 1 Pseudo-code proposed method

Require: K

$T \leftarrow$ **LDA**(K);
for all $w \in W$ **do**
  for all $k \in K$ **do**
    P$(w|k)$;
  end for
end for
for $n \leftarrow 1; n == k; n++$ **do**
  for $i \leftarrow 1; i == n; i++$ **do**
    $T_n \leftarrow$ **P**$(T_n + k_i)$;
  end for
end for
for all $n \in K$ **do**
  P$(w|Tn) \leftarrow \sum_{k=1}^{n} P(w|k)$;
  PL $\leftarrow ln\left(\frac{nw}{\sum_{w=1}^{nw}\frac{1}{P(w|T_n)}}\right)$;
end for
return MAXPL(T_n);

4 RESULTS AND DISCUSSION

To test de efficiency and efficacy of the method proposal, we present the application of the metric to two databases. We also apply metrics proposed in the literature primarily to compare processing times. We used two datasets of corpora of English texts (publicly available). Experiments were run on a 12 GB RAM computer.

The first dataset used comprised articles from The New York Times (*https* : *www.nytimes.com*)an American newspaper based in New York City with worldwide influence and readership. The second dataset used comprised Amazon customer reviews camera product (*https* : *s*3.*amazonaws.com*). The datasets are described in table 6.

Table 6 – Datasets descriptions.

-	New York Times	Customer reviews
Number of documents	42139	25000
Words	41162	20510
Period	June - December 2016	2015

The first step started with text pre-processing, this step refers to a data cleaning process. Filtering and cleaning the data by removing special characters and accents besides the removal of stopwords, to enhance quality and potentialize equally analysis. The development of the experiments was processed in R (R Core Team, 2020) with packages used in text mining and topic modeling. The code with the pre-processing, functions and packages used is available in Appendix A.

The metrics found in the literature were classified as punctual or graphic decisions (see table 5). The proposed metric has a characteristic of the decision based on "point", i.e., the result presented by the metric is a point value. To perform a better comparison between the metrics, only metrics with a punctual decision were considered.

In the two databases, we compute five metrics Griffiths and Steyvers (2004) [**Griffths**]; Cao et al. (2009) [**CaoJuan**]; Arun et al. (2010) [**Arun**] ; Deveaud et al. (2014) [**Deveaud**] and the metric proposed in this work [**Proposal**] for different K ranges: 2 to 25, 2 to 50, 2 to 75, 2 to 100 and 2 to 200. In the experiments, the Gibbs sampling algorithm with 1000 repetitions and burn-in of 500 repetitions was used. For a better visualization of the results, we normalized (scaled) all the values of the results of the metrics for the interval [0,1].

4.1 New York Times articles dataset

Our first experiment was carried out on the entire corpus described above, and the results are shown in table 7.

Table 7 – Experimental results with different metrics and different K ranges - NYT articles dataset.

METRICS	TOPICS				
	2 to 25	2 to 50	2 to 75	2 to 100	2 to 200
Griffths	12	4	8	9	5
CaoJuan	11	15	11	12	8
Arun	11	11	12	16	179
Deveaud	5	4	5	5	2
Proposal	2	2	4	6	4

We observed there are differences between the K intervals of different sizes. The metrics that showed less variation in the different K intervals are **CaoJuan**, **Deveaud** and **Proposal**. In **Arun**, although there is little variation in the best K in the initial intervals, when processed in the range 2 to 200, the value of the best K is much higher than the other intervals and other metrics as well. It is possible to observe that as the range of the interval increases, the values of the best K obtained with **Griffths** and **Proposal** become more similar.

To better understand the performance of the metrics, figure 10 shows the results of the experiment in the range of 2 to 100 for the different metrics. Appendix B shows the other graphs for different ranges of K. In **CaoJuan** and **Deveaud**, it is possible to observe a stabilization, which is indicative that the value of the best K will not vary as the interval increases. The graphs of **Proposal** and **Griffths**, despite showing slight fluctuations, indicate there will be no major changes in the best K as the interval increases. Only **Arun** is not informative in this situation.

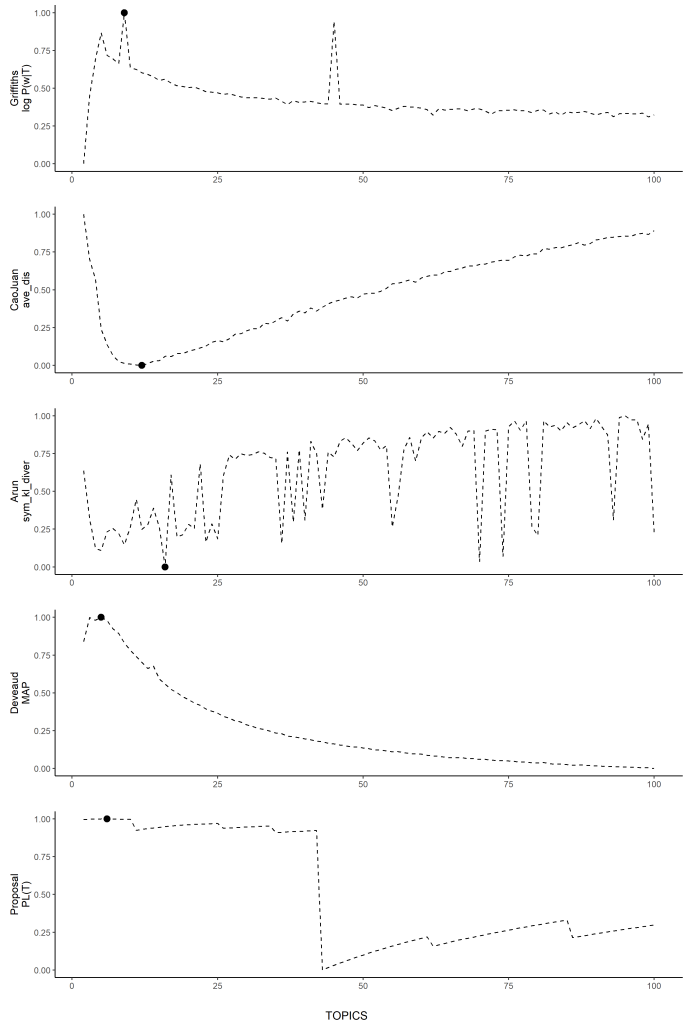


Figure 10 – Experimental results of the behavior of the different metrics - NYT articles dataset (input K of 2 to 100).

Word clouds are a widely used tool when doing text analysis. In the LDA model, word clouds are a way to visualize the probability that the word belongs to certain topics. The probabilistic weights of the words correspond to the graphic sizes (fonts) of the words, that is, the larger the font the more likely the word belongs to the topic.

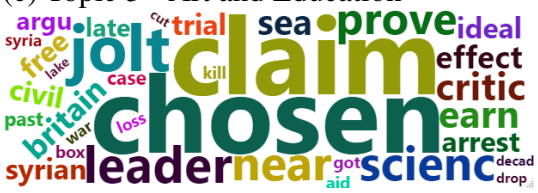
The highest best K value indicated using the **Proposal** metric was obtained with a K input in the range of 2 to 100,where the result divides the first dataset into 6 topics. Figure 11 gives word clouds for the model with 6 topics.

(a) Topic 1 - Sports

(b) Topic 2 - Politics and Elections

(c) Topic 3 - Art and Education

(d) Topic 4 - Business and Economy



(e) Topic 5 - International

(f) Topic 6 - New York City

Figure 11 – Word clouds with the 100 most relevant words in each topic - NYT articles dataset - **Proposal** [2 to 100].

Each of the six topic word clouds in figure 11 depict a unique and distinguishable theme, which correspond to distinct sections of interest typical of newspaper articles. Topic 1 was classified as issues related to Sports and Topic 2 as Politics and Elections. The database period corresponds to two important events, the 2016 Summer Olympics and 2016 United States presidential election, these events are evident in the word clouds (figure 11 (a) and figure 11 (b)).

Topic 3 labeled as Art and Education according to the words: book, novel, famous, photograph and television. Topic 4 was classified as issues related to Business and Economy according to the words: desing, job, recal and accord. Topic 5 was classified as issues related to International, although there are words that refer to Business, Economy and Politics, the words Islam, Berlin and diplomat indicative of their relationship with international affairs. Topic 6 was classified as issues related to New York City, but it is not as clear and evident as the others, it obtained this confirmation with the following analysis.

The word clouds obtained with the other values of K of the other metrics in the experiments in the range of 2 to 100 are illustrated in Appendix C.

Analyzing the word clouds of other values of K, it happens precisely the problems pointed out when choosing a large K where it is difficult to label the topic and classify them so it makes sense to a human (**Arun** 16 topics - figure 25 and **CaoJuan** 12 topics - figure 24).

Table 8 – Article title by topics - NYT articles dataset.

Topic	Article title	Probability
T1	Boston University Wont Celebrate Yet (the Season Just Started)	0.271
	Raiders Move Up, Chargers Move Down and Both May Be Moving Out	0.248
	Rio Olympics Today: Simone Biles Soars to Fourth Gold Medal	0.243
	Rio Olympics: Ashton Eaton and Usain Bolt Defend Titles	0.240
	Still Questioning the Best-of-Five Format in Mens Tennis	0.240
	U.S. Antidoping Agency Seeks to Depose Doctor Who Treated Top Track Athletes	0.237
	Jurgen Klinsmann Fired as U.S. Soccer Coach	0.236
	No Joy in Football? N.F.L. Celebration Penalties Rise Sharply	0.233
	At Yale, Theo Epstein Made a Splash as Newspapers Sports Editor	0.233
	A Paralympian Races to Remove Obstacles for the Next Generation	0.233
	Latino Players Rich History in Baseball, Now on Display at Smithsonian	0.230
T2	Is the Recent Spike in Marriages a Trump Bump?	0.259
	Smirk or Smile? The New York Tabloids Disagree About Trump	0.248
	Subway Sticky Notes Offer Post-Election Therapy	0.244
	Trump Won the Election, but 3 Manhattan Buildings Will Lose His Name	0.240
	New York Police on Alert After Warning of Terror Attack Before Election	0.240
	Seeking to Keep Control of New York Senate, G.O.P. Finds a Villain: Bill de Blasio	0.236
	Amid Division, a March in Washington Seeks to Bring Women Together	0.236
	Crowd Awaiting Clinton Is Stunned by Her Loss and Fearful of the Future	0.235
	In Rare New Jersey Swing District, House Race Becomes a Brawl	0.233
	Gomez Addams, Donald Trump and the Art of Failing	0.232
	Trumps Election? Some Students Are Too Busy to Worry	0.232
T3	For This Choreographer, Dance Speaks Truer Than Words	0.291
	After a Robust 2016, Jazzs Center Is Up for Grabs	0.280
	Vinyl Record Manufacturer in Nashville Is Said to Be Expanding	0.260
	In College Turmoil, Signs of a Changed Relationship With Students	0.259
	How Conservative Sites Turn Celebrity Despair on Its Head	0.252
	The Best TV Shows and Movies New to Netflix, Hulu and More in October	0.252
	Wall Street Dealmaker Says Professor Took Him for a Ride	0.252
	Thinking Critically About How We Engage With News Events Online and in Social Media	0.248
	Is Pokémon Go a Positive Cultural Force? Or Is it Just Another Excuse for People to Stare at Their Phones?	0.244
	Trigger Warnings, Safe Spaces and Microaggressions: Discussing Questions of Freedom of Speech on Campus	0.244
	Teaching How to Write Psychology Abstracts While Exploring Themes in News Articles	0.241
T4	Gender Diversity on Corporate Boards	0.370
	For Trumps Wealthy Cabinet, Prospect of a Sweet Tax Break	0.347
	Labor Board Challenges Secrecy in Wall Street Contracts	0.342
	Theranos, Embattled Laboratory, Shifts to Medical Machines	0.339
	Airbnb for Business Travelers: More Wi-Fi, Fewer Hosts in Towels	0.239
	Wary Corporate Chiefs Keep an Ear Tuned to Trumps Messages	0.236
	Amazon Is Quietly Eliminating List Prices	0.235
	Elizabeth Holmes of Theranos Is Barred From Running Lab for 2 Years	0.232
	Elizabeth Holmes, Founder of Theranos, Falls From Highest Perch Off Forbes List	0.232
	Start-Ups Selling Seats on Private Jets Dont Always Make It	0.232
	Message to Workers Under Scrutiny: Cooperate or Get Fired	0.229
T5	Canada Today: A Surge in Interest, and Prices, for Torontos Film Scene	0.276
	U.N. Chief Presses to Unlock Mystery of Dag Hammarskjolds Death	0.260
	French Prime Minister Faults Times Article Giving Voice to Muslim Women	0.259
	Discord Over Snooping Muted by Security Fears	0.244
	Reporter Who Wrote of Military-Civilian Clash Says He Cant Leave Pakistan	0.244
	Fidel Castro, Cuban Revolutionary Who Defied U.S., Dies at 90	0.244
	U.N.s Syria Envoy Suggests Donald Trump Has Limited Window to Work With Russia	0.243
	Britains Supreme Court Hears Legal Challenge to Brexit	0.236
	How Russia Recruited Elite Hackers for Its Cyberwar	0.234
	Italys Constitutional Referendum: What You Need to Know	0.233
	Russia and the U.S. Election: What We Know and Dont Know	0.232
T6	Second Avenue Subway Ringing in New Year With Party	0.260
	Subway Rider Arrested on Lewdness Charges Had a Giveaway: A Team U.S.A. Tattoo	0.252
	In a Year of Crime News, Some Dark Deeds Yield Dead Ends	0.243
	Inmate Seeks New Trial in 1993 Killing, Saying Ex-Detective Pressured Witness	0.243
	New York City Wants to Move 16- and 17-Year-Olds From Rikers Jail to Bronx Center	0.243
	Lives Upended by Disputed Cuts in Home-Health Care for Disabled Patients	0.243
	National Security Agency Said to Use Manhattan Tower as Listening Post	0.243
	What New York Hoaxes Have Rivaled the One on the Scary Clowns?	0.243
	Panthers, the Guardians of Prospect Park, Didnt See Citi Bike Coming	0.243
	At Carnegie Deli in Manhattan, Just 3 Months of Pastramis to Go	0.243
	As New York Fights Zika Virus, Officials Turn Their Focus to Sex	0.243

In table 8, a series of article titles from the dataset are presented that have their highest probabilistic association in each of the six topics. Most of these articles were subjectively judged to belong to the labeled topics, emphasizing that the value of K determined with the method proposed in this dissertation.

This second analysis is important for a more detailed and assertive verification to validate the model. Especially where it is difficult to interpret and connect the word cloud to a topic. As in Topic 6 - New York City (figure 11 (f)), where after analyzing the probabilistic association of the articles, a better understanding is possible.

4.1.1 Processing time - NYT articles dataset

In figure 12, we show the processing time of the experiment. The best performance is by the **Proposal** metric. Performing the processing time of this metric is superior when compared to that of the other metrics analyzed. This result becomes more evident as the amplitude of the K interval increases. In **Arun**, there is no certainty of behavior, especially when analyzing figure 10, where it is evident that there is no sign of convergence.

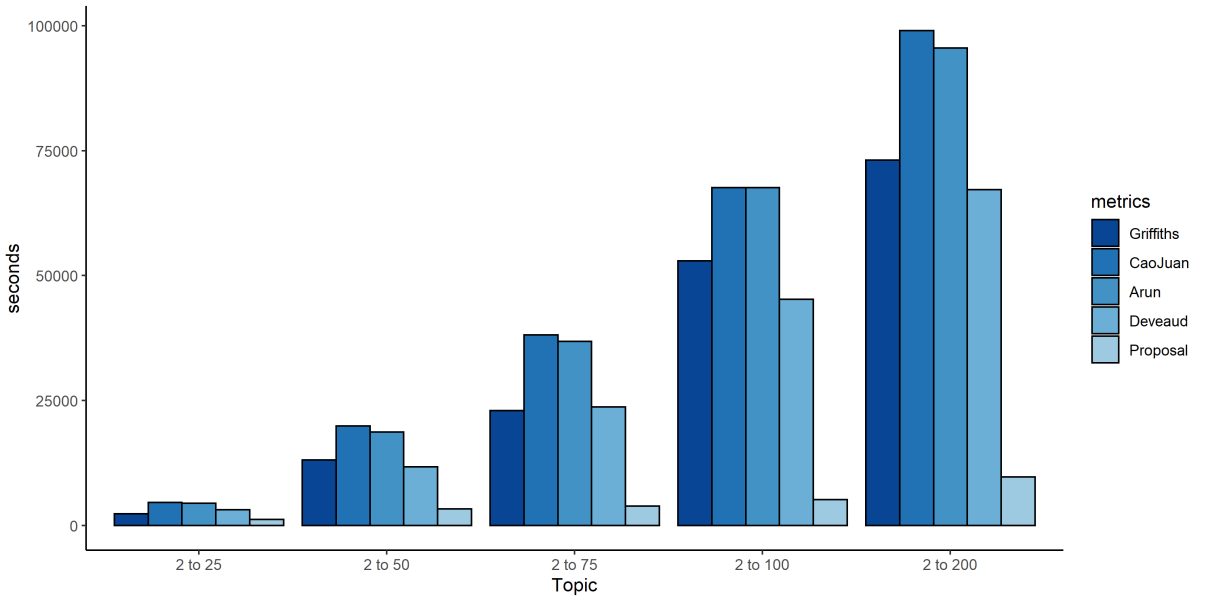


Figure 12 – Processing time - NYT articles dataset.

Figure 13 emphasizes the processing time and number of topics defined with input to find the value in the different methods.

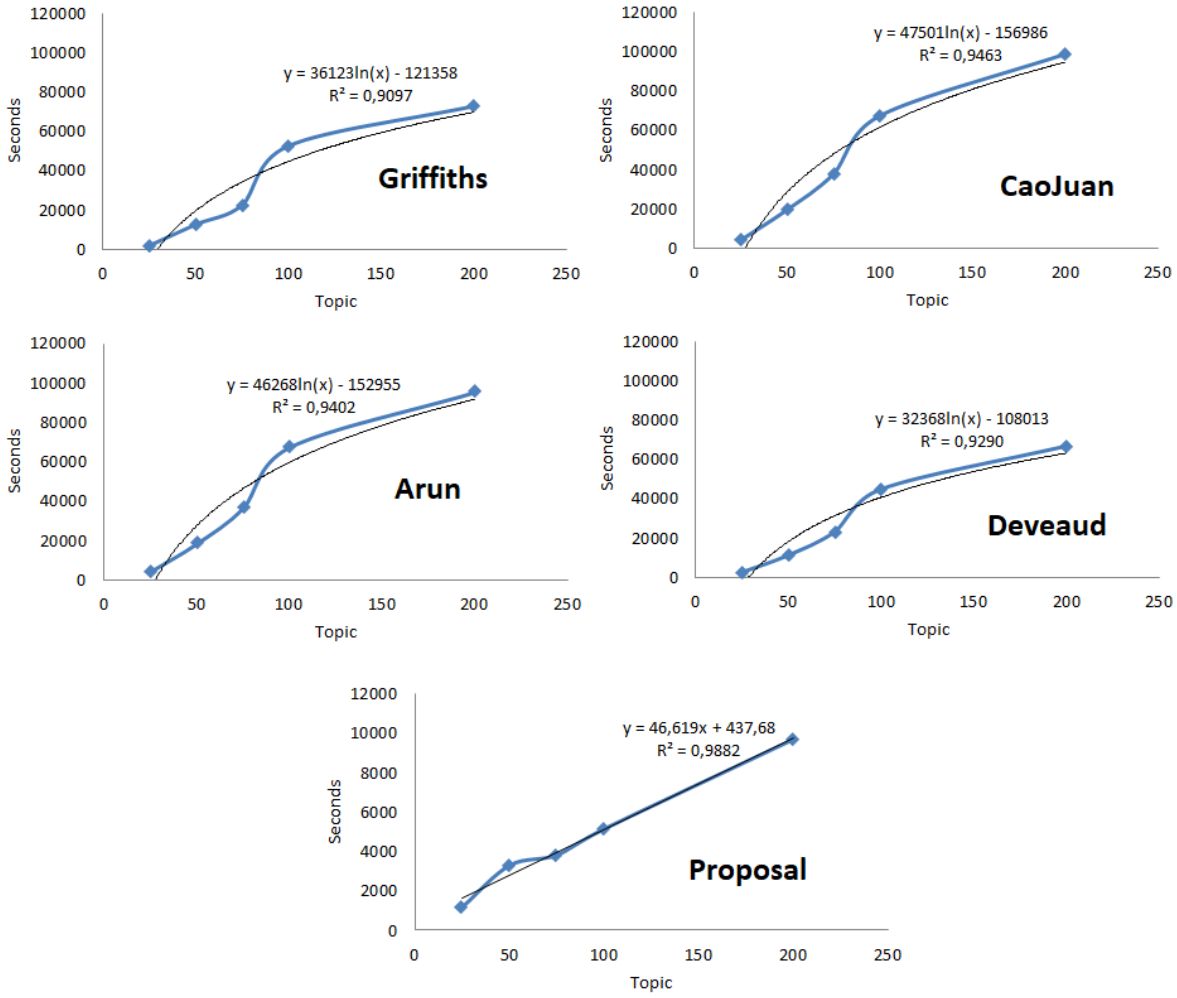


Figure 13 – Trend lines processing time versus number of topics - NYT articles dataset.

A logarithmic trendline is a best-fit curved line for the **Griffths**, **CaoJuan**, **Arun** and **Deveaud** methods for the database. With the **Proposal** method, a linear trend line was more appropriate.

4.2 Amazon customer reviews dataset

The second dataset used comprised Amazon customer reviews (camera product). The result of the second experiment carried out with the entire corpus described above is shown in table 9.

Table 9 – Experimental results with different metrics and different K ranges - Customer reviews dataset.

METRICS	TOPICS				
	2 to 25	2 to 50	2 to 75	2 to 100	2 to 200
Griffths	24	7	56	52	99
CaoJuan	3	4	3	4	3
Arun	25	22	23	78	117
Deveaud	2	2	2	2	2
Proposal	2	8	6	9	12

The **Arun** metric again shows a discrepancy in the results despite the **Griffths** metric tracking its behavior. **Deveaud** remains unchanged at different intervals.

In figure 14, we can see that the behavior of the metrics is identical to the first experiment, **Griffths** presents the peak, and after, despite oscillations, it does not reach the same level as the peak. **CaoJuan** and **Deveaud** reach the minimum and maximum, respectively, with a very low number of topics. As in the first, **Proposal** presents a structural break. Again **Arun** is not informative in this situation, as it has a lot of fluctuation. The highest best K value indicated using the **Proposal** metric was obtained with a K input in the range of 2 to 100, where the result divides the first dataset into 9 topics.

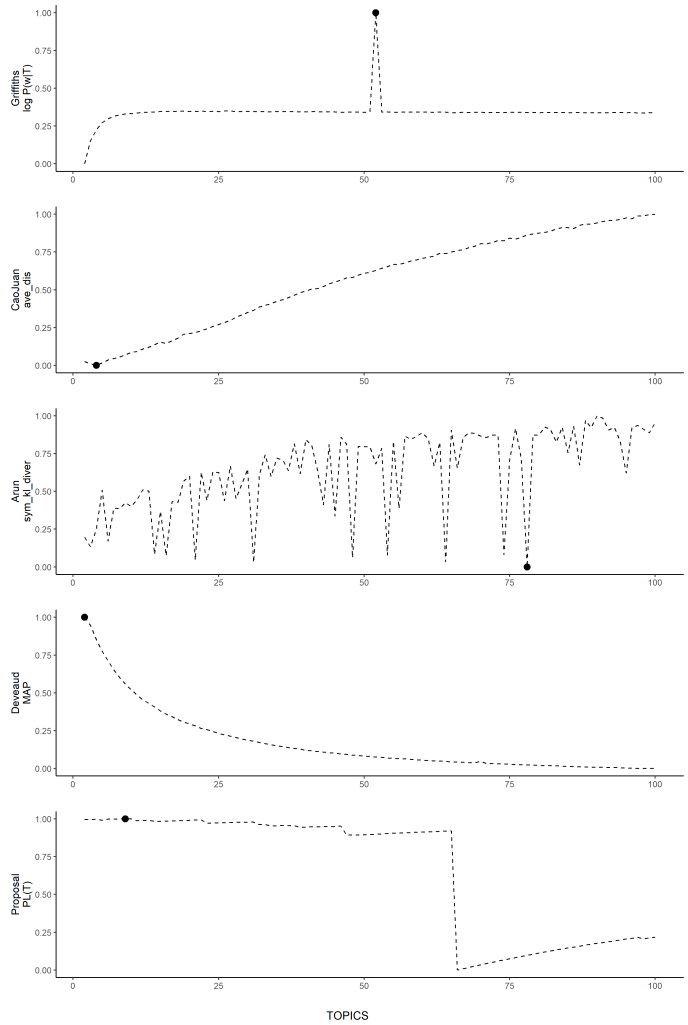
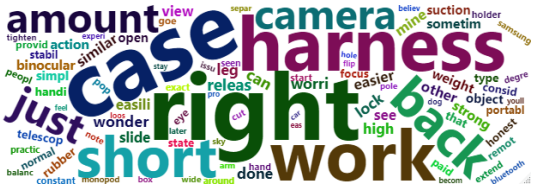


Figure 14 – Experimental results of the behavior of the different metrics - Customer reviews dataset (input K of 2 to 100)

The customer review dataset has no label available to classify them so it makes it possible to compare the construction of a cluster; a more qualitative approach is needed to assess and interpret the clusters. Figure 15 gives word clouds for the model with 9 topics.

(a) Topic 1 - Positive ratings



(b) Topic 2 - Accessories

(c) Topic 3 - Services

(d) Topic 4 - Prices

(e) Topic 5 - Pictures

(f) Topic 6 - Accessories

(g) Topic 7 - Functionality

(h) Topic 8 - Negative ratings

(i) Topic 9 - Accessories

Figure 15 – Word clouds with the 100 most relevant words in each topic - Customer reviews dataset - **Proposal** [2 to 100].

Topic 1 was classified as related to Positive ratings mainly by the words: great, well and easy. Topic 8 words with disappointed and very lead the cluster to be classified as Negative ratings. Topics 2, 6 and 9 were classified as Accessories according to the words: telescope and accessories (Topic 2); battery (Topic 6) and machine and card (Topic 9).

The words support and service indicate that Topic 3 is related to Services. Topic 4 was classified as questions related to Prices according to the words: price and value. Topic 5 was classified as issues related to Pictures, but it is not as clear and evident as the others, obtained this classification due to the word picture and work. Topic 7 is also not so clear and evident, it was classified as Functionality due to the words: function and camera.

The word clouds obtained with the other values of K of the other metrics in the experiments in the range of 2 to 100 are illustrated in Appendix D.

Table 10 – Customer reviews by topics - Customer reviews dataset.

Topic	Customer reviews	Probability
T1	Received item very quickly.....its works great!!!!	0.412
	Such a great camera for beginners especially! Easy to use while still producing quality photos and videos.	0.396
	Great product!!!! Incredible at evening out light.... Great for on the go photographers.	0.393
	Fantastic...Easy to use. Well constructed.	0.393
	Exceeded our expectations. Work great and I got these 4 for the same price of just ONE at Best buy.	0.387
	Great items to have.	0.387
T2	Battery life is short.	0.423
	Great back with just the right amount of compartments for my gear.	0.361
	Very good on batteries, recommended.	0.343
	Camera excellent some of the accessories are of questionable quality but time will tell.	0.327
	The camera harness is exceptional! I can hike, hands free and have my camera at the instant ready. Many thanks.	0.321
	Works as advertised. I look at the sun daily, if its out. Only drawback was that the telescope fits very tightly in the case.	0.304
T3	Excellent and excellent customer service. Highly recommended.	0.373
	Great customer service and product is made out of superior materials.	0.358
	Henry at tech support was great in setting up the camera. Awesome customer service.	0.355
	Amazing camera (love it!!), fast delivery and great customer service!	0.326
	Perfect, great quality at a rock bottom price! The customer service was Awesome too.	0.313
	Customer support is fabulous. Worked on my issue and continued until problem was fixed.	0.312
T4	Excellent quality price ratio.	0.397
	Great value. Cheaper then the stories.	0.355
	Quality exceeded expectations that were based on price.	0.350
	Excellent ball head, especially at that price!	0.330
	Much appreciated, if sold at more reasonable price.	0.321
	Excellent product for and excellent price!	0.296
T5	Works great, much better pic than the cheap one I had before, I need to get the second one installed out back real soon.	0.290
	Clear and crisp photos. Good price!	0.251
	Very easy to use and takes terrific pictures.	0.240
	It takes great pictures - I am quite pleased with it.	0.234
	Very nice night vision picture quality for the price.	0.234
	Great camera and the pictures come out awesome.	0.232
T6	It worked great. My battery was charged in no time.	0.372
	Battery is good.	0.328
	Much more power than battery that came with camera.	0.300
	Best extended battery than the gopro one.	0.299
	Battery arrived in time and is working beautifully.	0.297
	The battery charger works well and SterlingTek was terrific to work with.	0.281
T7	Just used it on our river trip in Europe and found the settings to be spot on. Threads easily and adjusts smoothly. A good universal fader.	0.304
	Absolutely love this camera. It has a lot of cool features!!!	0.268
	Great looking camera but did not check out the functions since I had returned the camera.	0.266
	Worked grate for what i was using it for.	0.260
	Works as advertised. All Olympus camera functions work.	0.258
	Very good device. A significant improvement in functionality over those available a couple of years ago.	0.254
T8	Very bad quality. In a two days one motor died. Bought new and replace. In a week another one died.	0.341
	Did not work seems like not charged even after plugging in.	0.329
	Terrible! Barely fits and doesnt show good quality. Very disappointed.	0.327
	This is crap. It was really hard to seat on my 5D Mark III and not accurate according to two other levels I tested it against. Is it possible to give this zero stars?	0.312
	Complete scam, fn cons, false advertising,...get a rope!	0.290
	If you want HD quality, dont get this. Sound wasnt great. Bought for a wedding and then a rock show, and was not impressed. Returning ASAP.	0.276
T9	Looks good but could not find 8gb sd card at all.	0.688
	Can not loop record and destroy 3 of my SD card.	0.289
	Very good credit card machine works well!	0.236
	Didnt come with the 8GB card and had to send back.	0.221
	Arrived very quick. I love how pictures are capture.	0.212
	In love with my camera brings 2 lenses which makes perfect.	0.210

In table 10, a series of Customer reviews of the dataset is presented, reviews with their highest probabilistic association in each of the nine topics.

This second analysis is possible to verify that the Topics 5 and 7 that presented more difficulties of interpretation have a low probability associated with Reviews in relation to other topics.

4.2.1 Processing time - Amazon customer reviews dataset

The processing time is somewhat limiting because in the **Griffths** metric, there is an indication it is necessary to increase the K amplitude to perform more tests, as well as in **Arun**, where it is evident that there is no sign of convergence, especially when analyzing the graphs (figure 10 and figure 14). In figure 16, it is evident that the processing time of the Proposal metric is better than the other metrics.

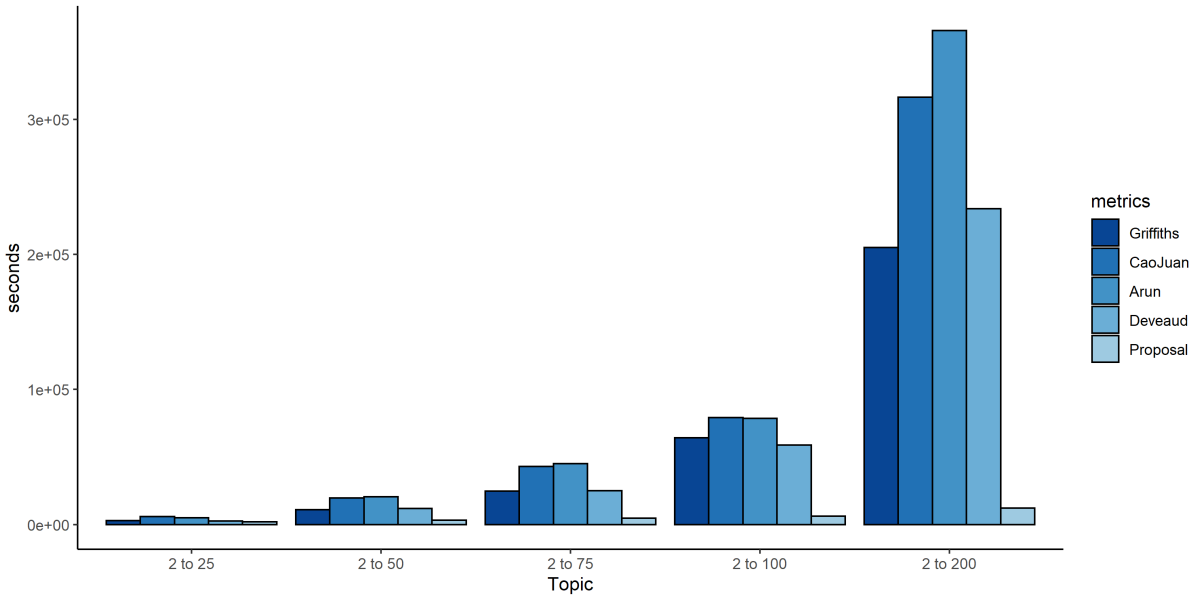


Figure 16 – Processing time - Customer reviews dataset.

In figure 17, it is possible to view the processing time and number of topics defined with input to find the value in the different methods in the Amazon customer reviews dataset.

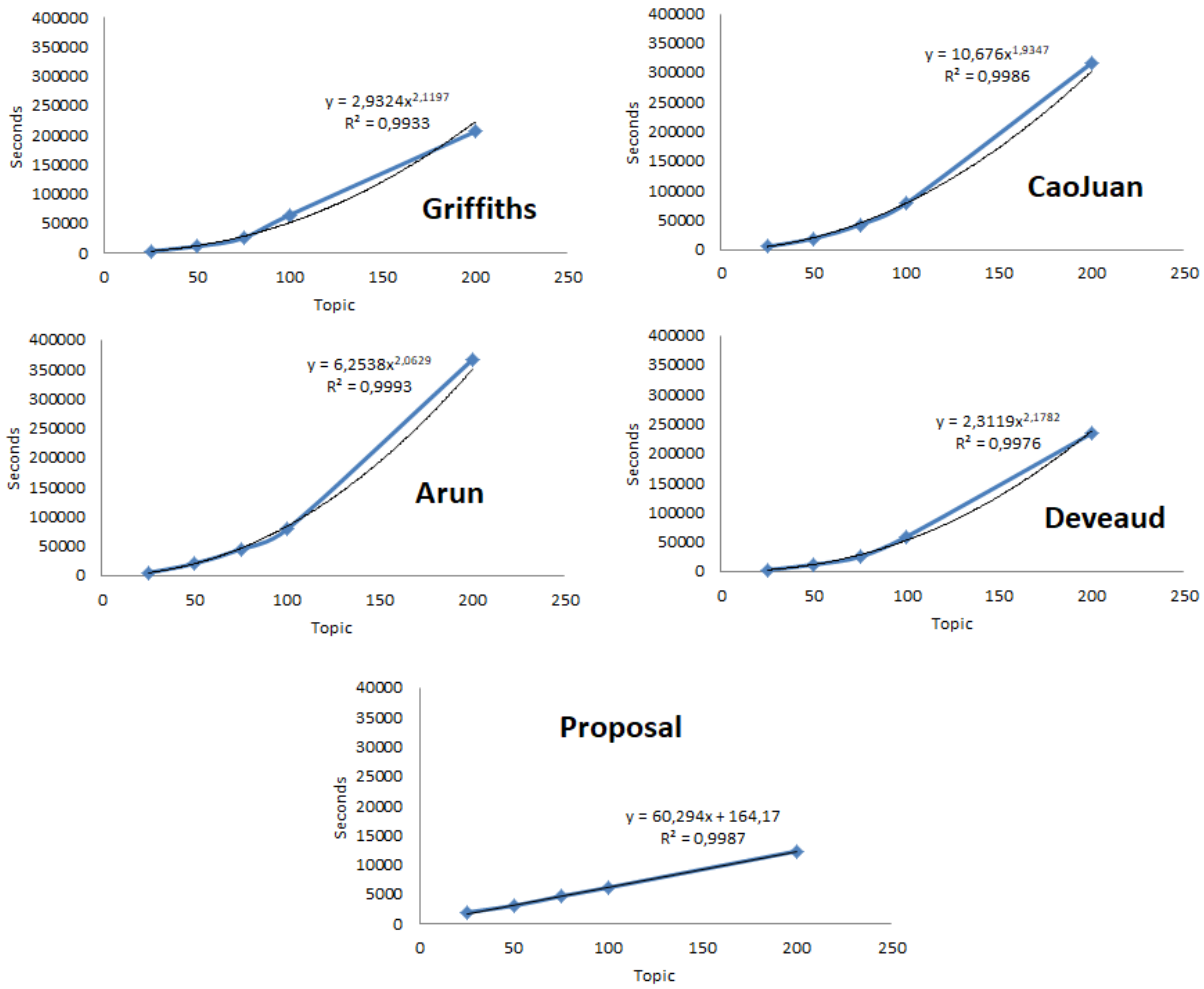


Figure 17 – Trend lines processing time versus number of topics - Customer reviews dataset.

A power trend line is a best fit curved line for the **Griffths**, **CaoJuan**, **Arun** and **Deveaud** methods for the database. The power trend line demonstrates the increase in time when the number of topics increases. With the **Proposal** method, again a linear trend line was more appropriate.

4.3 Discussion

Increasingly, researchers from different fields are using the LDA model, as evidenced in the item 2.2, a method borrowed from computer science. According to Zhao et al. (2015) topic modeling is a highly effective tool for text mining and knowledge discovery, entertaining requires skill and experience to apply successfully.

The text mining approach can be difficult to validate, tedious and often subjective, requiring a priori knowledge of the corpus by the researcherthe "best model" far beyond being statistically coherent needs to make sense in the real world. Liu et al. (2020) points out that an intractable limitation for LDA and its variants is that low-quality topics can be generated and

their meanings are confusing.

This dissertation explored the problem of selecting the parameter K of the LDA model for text analysis. We performed two experiments with different databases and compared their performance with other metrics found in the literature. In summary, the results presented previously confirm and reinforce the good performance of the metric proposed in this dissertation, mainly in the agility of the computational procedure. The metrics that were compared with the presented proposal are reasonable, but they carry a large load of time as evidenced previously (figure 12 and figure 16).

In Griffiths and Steyvers (2004), a behavior very similar to the first experiment is observed, where there is an oscillation when changing the input range, as the metric indicates there is this characteristic behavior indicating that with smaller intervals, the metric does not reach the so-called peak.

In both experiments, the Cao et al. (2009) metric remains stable and how he considers the correlation between topics is indicative there is little difference in correlation when the number of K increases.

The Arun et al. (2010) metric presents a significant discrepancy when the ranges amplitude increases. The metric considers word, topic and document information. The ranges amplitude considered influences the construction of the matrices and affects the results.

In Deveaud et al. (2014) remains practically unchanged in the different intervals in both datasets. Its objective is to find the most dispersed model, which indicates that when considering the model with few topics, this characteristic is found.

The Proposed metric presents a characteristic behavior of an abrupt drop after reaching the maximum peak and suggests an intermediate number of K considering the suggestion of the other metrics. Although each metric has its own method for determining the number of topics, some results are similar for the same database, as evidenced in the results.

The first dataset (New York Times) presents more words (42139 words) compared to the second dataset (Customer reviews) where there are 41162 words, analyzing the results of better K in the metrics in the experiments the second dataset more topics than the first dataset . This result corroborates with the Sbalchiero-Eder rule (SBALCHIERO; EDER, 2020) that the larger the portions of the text, the smaller, the better number of topics, emphasizing that the relationship is not linear.

The main limitation of our approach is that we do not consider the construction of several analysis models like the other four metrics; this is a limiting factor in our metric. However, experiments show that their solutions do not differ from those determined with the parameters of others.

5 CONSIDERATIONS AND FUTURE WORK

In this dissertation, we propose a metric to determine the number of topics in the LDA model. Determining the number of parameter topics is extremely important and little explored in the literature, mainly due to the lengthy computation procedure, which makes researchers choose the number of topics arbitrarily. The problem solved consists of subsidizing the decision regarding the number of topics with a shorter processing time. The goal of this dissertation is, therefore, to propose a metric to determine the number of topics in the LDA model and to compare its performance with that of other metrics, mainly considering the processing time. In the LDA model, the number of topics is a fundamental parameter for the construction of the model, as it is reflected in the data analysis.

Our first specific objective leads us to identify existing metrics to determine the number of the parameter K in the LDA model, we came to classify the metrics in the form of selection of the number of topics (graphical and point analysis), which led us to our second and third specific objectives, where we implemented the metrics with punctual selection in two corpus and compared the performance and the processing time.

Although each metric has its own method for determining the number of topics, some results are similar for the same database, as evidenced in the study. Our metric is superior when considering processing time. Experiments show this method is effective.

George and Doss (2017) points out that when making the selection of models in modeling textual topics, there are two competing objectives. One would be to select the correct model statistically and the other objective is to select the model that provides the best inference, that is, data interpretation.

The most likely words presented for each topic are semantically associated, they are not only discovered by the LDA model (modeling of unsupervised topics), but they can be perfectly grasped by human judgment.

Topic modeling has gained ground in recent years in research in different areas and the LDA model represents one of the most widespread topic modeling algorithms, considered a milestone in the topic models panorama (SBALCHIERO; EDER, 2020; BLAIR et al., 2020).

In one of our surveys, we assessed the classification of open responses in a customer satisfaction survey using the LDA model. Comparing the classification of customer responses using the LDA model with the classification of three researchers (800 questionnaires with three questions each). The accuracy of the classification was over 80%, the LDA technique classified the attributes and emerging words, highlighting problems to be solved by supermarket man-

agers. Another highlight of the study was the processing time, the time (classification only) to perform the classification via LDA was approximately 1 minute, and the time to perform the classification task by researchers 1, 2 and 3 was 18h 32 min, 14h 32 min and 15:21 min respectively (LIMA Jr; BECKER, 2020b).

In another of our surveys, the result of the presented dissertation project, we investigated the relationship between the Ibovespa index and the themes extracted from the New York Times (textual data). We evaluated the long-term impacts when a variable receives a shock. The news from the New York Times (January 2014 to December 2016) was organized daily. With the LDA technique, it was possible to identify emerging topics, and thus it was possible to work with numerical and textual data. It was possible to observe the absence of two-dimensional relationships between the variables. With the results of the Granger causality test, the relationship between the Ibovespa Index and Topic 2 (New York Times) was identified, which was labeled as Elections and presidential campaign (LIMA Jr; BECKER, 2020a).

The results obtained with the proposal of a new metric to determine the number of topics in the LDA model presented here open, in our opinion, new horizons for comparative research in other data sets. A new possibility arises that deserves attention, the investigation of the behavior characterized by an abrupt fall after reaching the maximum peak, indicating a structural break.

Bibliography

AGARWAL, R. P. Pythagorean theorem before and after pythagoras. **Adv. Stud. Contemp. Math**, v. 30, 2020.

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. : Springer Science & Business Media, 2012.

AL-OBEIDAT, F.; SPENCER, B.; KAFEZA, E. The opinion management framework: Identifying and addressing customer concerns extracted from online product reviews. **Electronic Commerce Research and Applications**, v. 27, p. 52 – 64, 2018. ISSN 1567-4223. Available in: <http://www.sciencedirect.com/science/article/pii/S1567422317300923>.

ALBUQUERQUE, P. H.; VALLE, D. R. do; LI, D. Bayesian lda for mixed-membership clustering analysis: The rlda package. **Knowledge-Based Systems**, Elsevier, v. 163, p. 988–995, 2019.

ALP, Z. Z.; ÖðüDüCü, S. G. Extracting topical information of tweets using hashtags. In: **2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)**. 2015. p. 644–648.

ARAúJO, V. Mecanismos de alinhamento de preferências em governos multipartidários: controle de políticas públicas no presidencialismo brasileiro. **Opinião Pública**, scielo, v. 23, p. 429 – 458, 08 2017. ISSN 0104-6276. Available in: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-62762017000200429&nrm=iso>.

ARORA, S.; GE, R.; MOITRA, A. Learning topic models–going beyond svd. In: IEEE. **Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on**. 2012. p. 1–10.

ARUN, K. S.; GOVINDAN, V. K. A hybrid deep learning architecture for latent topic-based image retrieval. **Data Science and Engineering**, v. 3, n. 2, p. 166–195, Jun 2018. ISSN 2364-1541. Available in: <https://doi.org/10.1007/s41019-018-0063-7>.

ARUN, R.; SURESH, V.; MADHAVAN, C. E. V.; MURTHY, M. N. N. On finding the natural number of topics with latent dirichlet allocation: Some observations. In: ZAKI, M. J.; YU, J. X.; RAVINDRAN, B.; PUDI, V. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 391–402. ISBN 978-3-642-13657-3.

ASNANI, K.; PAWAR, J. D. Automatic aspect extraction using lexical semantic knowledge in code-mixed context. **Procedia Computer Science**, v. 112, p. 693 – 702, 2017. ISSN 1877-0509. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France. Available in: <http://www.sciencedirect.com/science/article/pii/S187705091731503X>.

AUGUIE, B.; ANTONOV, A.; AUGUIE, M. B. Package 'gridextra'. r package version 2.3. **Miscellaneous Functions for Grid Graphics**, 2017.

BAECHLE, C.; HUANG, C. D.; AGARWAL, A.; BEHARA, R. S.; GOO, J. Latent topic ensemble learning for hospital readmission cost optimization. **European Journal of Operational Research**, v. 281, n. 3, p. 517 – 531, 2020. ISSN 0377-2217. Featured Cluster:

Business Analytics: Defining the field and identifying a research agenda. Available in: <http://www.sciencedirect.com/science/article/pii/S0377221719304102>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. et al. **Modern information retrieval**. : ACM press New York, 1999. v. 463.

BALAKRISHNAN, V.; LOK, P. Y.; RAHIM, H. A. A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews. **The Journal of Supercomputing**, Springer, p. 1–16, 2020.

BAO, Y.; DATTA, A. Simultaneously discovering and quantifying risk types from textual risk disclosures. **Management Science**, v. 60, n. 6, p. 1371–1391, 2014. Available in: <https://doi.org/10.1287/mnsc.2014.1930>.

BASTANI, K.; NAMAVARI, H.; SHAFFER, J. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. **Expert Systems with Applications**, Elsevier, v. 127, p. 256–271, 2019.

BELLOGíN, A.; CANTADOR, I.; CASTELLS, P. A comparative study of heterogeneous item recommendations in social systems. **Information Sciences**, v. 221, p. 142 – 169, 2013. ISSN 0020-0255. Available in: <http://www.sciencedirect.com/science/article/pii/S0020025512006329>.

BIMBA, A. T.; IDRIS, N.; AL-HUNAIYYAN, A.; MAHMUD, R. B.; ABDELAZIZ, A.; KHAN, S.; CHANG, V. Towards knowledge modeling and manipulation technologies: A survey. **International Journal of Information Management**, v. 36, n. 6, Part A, p. 857 – 871, 2016. ISSN 0268-4012. Available in: <http://www.sciencedirect.com/science/article/pii/S026840121630336X>.

BISHOP, C. M. **Pattern recognition and machine learning**. : springer, 2006.

BLAIR, S. J.; BI, Y.; MULVENNA, M. D. Aggregated topic models for increasing social media topic coherence. **Applied Intelligence**, Springer, v. 50, n. 1, p. 138–156, 2020.

BLAZQUEZ, D.; DOMENECH, J. Big data sources and methods for social and economic analyses. **Technological Forecasting and Social Change**, v. 130, p. 99 – 113, 2018. ISSN 0040-1625. Available in: <http://www.sciencedirect.com/science/article/pii/S0040162517310946>.

BLEI, D. M. Probabilistic topic models. **Commun. ACM**, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. ISSN 0001-0782. Available in: <http://doi.acm.org/10.1145/2133806.2133826>.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BOLELLI, L.; ERTEKIN, S.; ZHOU, D.; GILES, C. L. Finding topic trends in digital libraries. In: **Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries**. New York, NY, USA: ACM, 2009. (JCDL '09), p. 69–72. ISBN 978-1-60558-322-8. Available in: <http://doi.acm.org/10.1145/1555400.1555411>.

BOST, X.; SENAY, G.; EL-BèZE, M.; MORI, R. D. Multiple topic identification in human/human conversations. **Computer Speech & Language**, v. 34, n. 1, p. 18 – 42, 2015. ISSN 0885-2308. Available in: <http://www.sciencedirect.com/science/article/pii/S0885230815000352>.

BOUCHET-VALAT, M. Package 'snowballc: Snowball stemmers based on the c libstemmer utf-8 library'. r package version 0.7.0. **R package**, v. 1, 2020.

BOYER, C. B.; MERZBACH, U. C. **A history of mathematics**. : John Wiley Sons, 2011.

BROWN, N. C.; CROWLEY, R. M.; ELLIOTT, W. B. What are you saying? using topic to detect financial misreporting. **Journal of Accounting Research**, v. 58, n. 1, p. 237–291, 2020. Available in: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12294>.

BRYCHCíN, T.; KONOPíK, M. Semantic spaces for improving language modeling. **Computer Speech & Language**, v. 28, n. 1, p. 192 – 209, 2014. ISSN 0885-2308. Available in: <http://www.sciencedirect.com/science/article/pii/S0885230813000387>.

CAHYANINGTYAS, R. M.; KUSUMANINGRUM, R.; SUTIKNO; SUHARTONO; RIYANTO, D. E. Emotion detection of tweets in indonesian language using lda and expression symbol conversion. In: **2017 1st International Conference on Informatics and Computational Sciences (ICICoS)**. 2017. p. 253–258.

CALOMIRIS, C. W.; MAMAYSKY, H. How news and its context drive risk and returns around the world. **Journal of Financial Economics**, v. 133, n. 2, p. 299 – 336, 2019. ISSN 0304-405X. Available in: <http://www.sciencedirect.com/science/article/pii/S0304405X18303180>.

CAO, J.; XIA, T.; LI, J.; ZHANG, Y.; TANG, S. A density-based method for adaptive lda model selection. **Neurocomputing**, v. 72, n. 7, p. 1775 – 1781, 2009. ISSN 0925-2312. Advances in Machine Learning and Computational Intelligence. Available in: <http://www.sciencedirect.com/science/article/pii/S092523120800372X>.

CERISARA, C. Automatic discovery of topics and acoustic morphemes from speech. **Computer Speech & Language**, v. 23, n. 2, p. 220 – 239, 2009. ISSN 0885-2308. Available in: <http://www.sciencedirect.com/science/article/pii/S0885230808000387>.

CHANG, J.; GERRISH, S.; WANG, C.; BOYD-GRABER, J.; BLEI, D. Reading tea leaves: How humans interpret topic models. **Advances in neural information processing systems**, v. 22, p. 288–296, 2009.

CHEN, C.; REN, J. Forum latent dirichlet allocation for user interest discovery. **Knowledge-Based Systems**, v. 126, p. 1 – 7, 2017. ISSN 0950-7051. Available in: <http://www.sciencedirect.com/science/article/pii/S0950705117301727>.

CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business intelligence and analytics: From big data to big impact. **MIS quarterly**, v. 36, n. 4, 2012.

CHEN, J.-Y.; ZHENG, H.-T.; JIANG, Y.; XIA, S.-T.; ZHAO, C.-Z. A probabilistic model for semantic advertising. **Knowledge and Information Systems**, Feb 2018. ISSN 0219-3116. Available in: <https://doi.org/10.1007/s10115-018-1160-7>.

CHEN, L.-C. An effective lda-based time topic model to improve blog search performance. **Information Processing Management**, v. 53, n. 6, p. 1299 – 1319, 2017. ISSN 0306-4573. Available in: <http://www.sciencedirect.com/science/article/pii/S0306457317300997>.

CHENG, D.; HE, X.; LIU, Y. Model selection for topic models via spectral decomposition. In: **Artificial Intelligence and Statistics**. 2015. p. 183–191.

CHRISTIDIS, K.; MENTZAS, G.; APOSTOLOU, D. Using latent topics to enhance search and recommendation in enterprise social software. **Expert Systems with Applications**, v. 39, n. 10, p. 9297 – 9307, 2012. ISSN 0957-4174. Available in: <http://www.sciencedirect.com/science/article/pii/S095741741200317X>.

CHUI, M.; MANYIKA, J.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; SARRAZIN, H.; SANDS, G.; WESTERGREN, M. **The social economy: Unlocking value and productivity through social technologies**. 2012. v. 4.

DAVENPORT, T. Beyond unicorns: Educating, classifying, and certifying business data scientists. **Harvard Data Science Review**, 5 2020. Https://hdsr.mitpress.mit.edu/pub/t37qjoi7. Available in: <https://hdsr.mitpress.mit.edu/pub/t37qjoi7>.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. **Journal of the American society for information science**, American Documentation Institute, v. 41, n. 6, p. 391, 1990.

DEVEAUD, R.; SANJUAN, E.; BELLOT, P. Accurate and effective latent concept modeling for ad hoc information retrieval. **Document numérique**, Lavoisier, Cachan, v. 17, n. 1, p. 61–84, 2014. ISSN 9782746246546. Available in: <https://www.cairn.info/revue-document-numerique-2014-1-page-61.htm>.

DONG, H.; HUSSAIN, F. K.; CHANG, E. A survey in traditional information retrieval models. In: IEEE. **Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on**. 2008. p. 397–402.

ESCALANTE, H. J.; MONTES, M.; SUCAR, L. E. Multi-class particle swarm model selection for automatic image annotation. **Expert Systems with Applications**, v. 39, n. 12, p. 11011 – 11021, 2012. ISSN 0957-4174. Available in: <http://www.sciencedirect.com/science/article/pii/S0957417412004939>.

FEI-FEI, L.; PERONA, P. A bayesian hierarchical model for learning natural scene categories. In: IEEE. **Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on**. 2005. v. 2, p. 524–531.

FEINERER, I.; HORNIK, K.; FEINERER, M. I. Package tm. r package version 0.7-7. **Corpus**, v. 10, n. 1, 2019.

FUJIMOTO, H.; ETOH, M.; KINNO, A.; AKINAGA, Y. Topic analysis of web user behavior using lda model on proxy logs. In: HUANG, J. Z.; CAO, L.; SRIVASTAVA, J. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 525–536. ISBN 978-3-642-20841-6.

FULEKY, P. **Macroeconomic Forecasting in the Era of Big Data: Theory and Practice**. : Springer, 2019. v. 52.

GAO, L.; YU, Y.; LIANG, W. Public transit customer satisfaction dimensions discovery from online reviews. **Urban Rail Transit**, Springer, v. 2, n. 3-4, p. 146–152, 2016.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, n. 6, p. 721–741, 1984.

GEORGE, C. P.; DOSS, H. Principled selection of hyperparameters in the latent dirichlet allocation model. **The Journal of Machine Learning Research**, JMLR. org, v. 18, n. 1, p. 5937–5974, 2017.

GHAVAMI, P. **Big Data Analytics Methods: Analytics Techniques in Data Mining, Deep Learning and Natural Language Processing**. : Walter de Gruyter GmbH & Co KG, 2019.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. **Proceedings of the National academy of Sciences**, National Acad Sciences, v. 101, n. suppl 1, p. 5228–5235, 2004.

GRIMALDI, M.; CORVELLO, V.; MAURO, A. D.; SCARMOZZINO, E. A systematic literature review on intangible assets and open innovation. **Knowledge Management Research & Practice**, v. 15, n. 1, p. 90–100, Feb 2017. ISSN 1477-8246. Available in: <https://doi.org/10.1057/s41275-016-0041-7>.

GRUN, B.; HORNIK, K.; GRUN, M. B. Package topicmodels. r package version 0.2-11. 2020.

HARDT, M. S. **Who, what and when: how media and politicians shape the Brazilian debate on foreign affairs**. Tese (Doutorado) — Universidade de São Paulo, 2019.

HE, J.; LIU, H.; XIONG, H. Socotraveler: Travel-package recommendations leveraging social influence of different relationship types. **Information & Management**, v. 53, n. 8, p. 934 – 950, 2016. ISSN 0378-7206. Big Data Commerce. Available in: <http://www.sciencedirect.com/science/article/pii/S0378720616300362>.

HE, W.; FANG, Y.; MALEKIAN, R.; LI, Z. Time Series Analysis of Online Public Opinions in Colleges and Universities and its Sustainability. **Sustainability**, v. 11, n. 13, p. 1–17, June 2019. Available in: <https://ideas.repec.org/a/gam/jsusta/v11y2019i13p3546-d243535.html>.

HINDLE, G.; KUNC, M.; MORTENSEN, M.; OZTEKIN, A.; VIDGEN, R. Business analytics: Defining the field and identifying a research agenda. **European Journal of Operational Research**, v. 281, n. 3, p. 483 – 490, 2020. ISSN 0377-2217. Featured Cluster: Business Analytics: Defining the field and identifying a research agenda. Available in: <http://www.sciencedirect.com/science/article/pii/S0377221719308173>.

HOFMANN, T. Probabilistic latent semantic indexing. In: ACM. **Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval**. 1999. p. 50–57.

HU, Y.-H.; CHEN, Y.-L.; CHOU, H.-L. Opinion mining from online hotel reviews a text summarization approach. **Information Processing & Management**, v. 53, n. 2, p. 436 – 449, 2017. ISSN 0306-4573. Available in: <http://www.sciencedirect.com/science/article/pii/S0306457316306781>.

HUANG, J.; KALBARCZYK, Z.; NICOL, D. M. Knowledge discovery from big data for intrusion detection using lda. In: **2014 IEEE International Congress on Big Data**. 2014. p. 760–761. ISSN 2379-7703.

JAYARATNE, M.; JAYATILLEKE, B. Predicting personality using answers to open-ended interview questions. **IEEE Access**, v. 8, p. 115345–115355, 2020.

JO, T. **Text mining: Concepts, implementation, and big data challenge**. : Springer, 2018. v. 45.

JO, Y.; OH, A. H. Aspect and sentiment unification model for online review analysis. In: ACM. **Proceedings of the fourth ACM international conference on Web search and data mining**. 2011. p. 815–824.

JUNG, Y.; SUH, Y. Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. **Decision Support Systems**, v. 123, p. 113074, 2019. ISSN 0167-9236. Available in: <http://www.sciencedirect.com/science/article/pii/S0167923619301034>.

Kanungsukkasem, N.; Leelanupab, T. Financial latent dirichlet allocation (finlda): Feature extraction in text and data mining for financial time series prediction. **IEEE Access**, v. 7, p. 71645–71664, 2019.

KAPPASSOV, Z.; CORRALES, J.-A.; PERDEREAU, V. Tactile sensing in dexterous robot hands review. **Robotics and Autonomous Systems**, v. 74, p. 195 – 220, 2015. ISSN 0921-8890. Available in: <http://www.sciencedirect.com/science/article/pii/S0921889015001621>.

KAR, A. K.; DWIVEDI, Y. K. Theory building with big data-driven research–moving away from the what towards the why. **International Journal of Information Management**, Elsevier, v. 54, p. 102205, 2020.

KASZUBOWSKI, E. **Modelo de tópicos para associações livres**. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2016.

KIM, D.-k.; MOTOYAMA, M.; VOELKER, G. M.; SAUL, L. K. Topic modeling of freelance job postings to monitor web service abuse. In: **Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence**. New York, NY, USA: ACM, 2011. (AISec '11), p. 11–20. ISBN 978-1-4503-1003-1. Available in: <http://doi.acm.org/10.1145/2046684.2046687>.

KIM, H. D.; PARK, D. H.; LU, Y.; ZHAI, C. Enriching text representation with frequent pattern mining for probabilistic topic modeling. **Proceedings of the American Society for Information Science and Technology**, Wiley Online Library, v. 49, n. 1, p. 1–10, 2012.

KIM, H.-N.; SADDIK, A. E. A stochastic approach to group recommendations in social media systems. **Information Systems**, v. 50, p. 76 – 93, 2015. ISSN 0306-4379. Available in: <http://www.sciencedirect.com/science/article/pii/S0306437914001537>.

LAI, C.-H.; HONG, P.-R. Group recommendation based on the analysis of group influence and review content. In: NGUYEN, N. T.; TOJO, S.; NGUYEN, L. M.; TRAWIŃSKI, B. (Ed.). **Intelligent Information and Database Systems**. Cham: Springer International Publishing, 2017. p. 100–109. ISBN 978-3-319-54472-4.

LAYMAN, L.; NIKORA, A. P.; MEEK, J.; MENZIES, T. Topic modeling of nasa space system problem reports: Research in practice. In: **2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)**. 2016. p. 303–314.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, Nature Publishing Group, v. 401, n. 6755, p. 788–791, 1999.

LIANG, Y.; XU, F.; ZHANG, S.-H.; LAI, Y.-K.; MU, T. Knowledge graph construction with structure and parameter learning for indoor scene design. **Computational Visual Media**, v. 4, n. 2, p. 123–137, Jun 2018. ISSN 2096-0662. Available in: <https://doi.org/10.1007/s41095-018-0110-3>.

LIAO, X.; CHEN, G.; KU, B.; NARULA, R.; DUNCAN, J. Text mining methods applied to insurance company customer calls: A case study. **North American Actuarial Journal**, Taylor & Francis, v. 24, n. 1, p. 153–163, 2020.

LIMA Jr, A. V.; BECKER, J. L. Analysis of interactions between variables using topic modeling. In: **Anais do LII Simpósio Brasileiro de Pesquisa Operacional**. 2020.

LIMA Jr, A. V.; BECKER, J. L. Evaluation of topic models in consumer research. In: **Conference Proceedings BALAS 2020**. 2020.

LIU, Y.; DU, F.; SUN, J.; JIANG, Y. ilda: An interactive latent dirichlet allocation model to improve topic quality. **Journal of Information Science**, SAGE Publications Sage UK: London, England, v. 46, n. 1, p. 23–40, 2020.

LIU, Y.; HUANG, X.; AN, A.; YU, X. Arsa: A sentiment-aware model for predicting sales performance using blogs. In: **Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2007. (SIGIR '07), p. 607–614. ISBN 978-1-59593-597-7. Available in: <http://doi.acm.org/10.1145/1277741.1277845>.

LIVNE, A.; SIMMONS, M. P.; ADAR, E.; ADAMIC, L. A. The party is over here: Structure and content in the 2010 election. **ICWSM**, v. 11, p. 17–21, 2011.

LU, K.; CAI, X.; AJIFERUKE, I.; WOLFRAM, D. Vocabulary size and its effect on topic representation. **Information Processing & Management**, Elsevier, v. 53, n. 3, p. 653–665, 2017.

LUCINI, F. R.; TONETTO, L. M.; FOGLIATTO, F. S.; ANZANELLO, M. J. Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. **Journal of Air Transport Management**, Elsevier, v. 83, p. 101760, 2020.

MA, T.; LI, J.; LIANG, X.; TIAN, Y.; AL-DHELAAN, A.; AL-DHELAAN, M. A time-series based aggregation scheme for topic detection in weibo short texts. **Physica A: Statistical Mechanics and its Applications**, v. 536, p. 120972, 2019. ISSN 0378-4371. Available in: <http://www.sciencedirect.com/science/article/pii/S037843711930576X>.

MA, Y.; CHEN, G.; WEI, Q. Finding users preferences from large-scale online reviews for personalized recommendation. **Electronic Commerce Research**, v. 17, n. 1, p. 3–29, Mar 2017. ISSN 1572-9362. Available in: <https://doi.org/10.1007/s10660-016-9240-9>.

MADDING, C.; ANSARI, A.; BALLENGER, C.; THOTA, A. Topic modeling to understand technology talent. **SMU Data Science Review**, v. 3, n. 2, p. 16, 2020.

MAKHABEL, B. **R: Mining Spatial, Text, Web, and Social Media Data: Create and Customize Data Mioning Algorithms: a Course in Three Modules**. : Packt Publishing, 2017.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. et al. **Introduction to information retrieval**. : Cambridge university press Cambridge, 2008. v. 1.

MARCOLIN, C.; BECKER, J. A. L.; WILD, F.; SCHIAVI, G.; BEHR, A. Business Analytics in Tourism: Uncovering Knowledge from Crowds. **BAR - Brazilian Administration Review**, scielo, v. 16, 00 2019. ISSN 1807-7692. Available in: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-76922019000200305&nrm=iso>.

MATSUTANI, T.; HAMADA, M. Parallelized latent dirichlet allocation provides a novel interpretability of mutation signatures in cancer genomes. **Genes**, Multidisciplinary Digital Publishing Institute, v. 11, n. 10, p. 1127, 2020.

MCCALLUM, A.; WANG, X.; CORRADA-EMMANUEL, A. Topic and role discovery in social networks with experiments on enron and academic email. **Journal of Artificial Intelligence Research**, v. 30, p. 249–272, 2007.

MCLAURIN, E. J.; LEE, J. D.; MCDONALD, A. D.; AKSAN, N.; DAWSON, J.; TIPPIN, J.; RIZZO, M. Using topic modeling to develop multi-level descriptions of naturalistic driving data from drivers with and without sleep apnea. **Transportation Research Part F: Traffic Psychology and Behaviour**, v. 58, p. 25 – 38, 2018. ISSN 1369-8478. Available in: <http://www.sciencedirect.com/science/article/pii/S1369847817302553>.

MIN, K.-B.; SONG, S.-H.; MIN, J.-Y. Topic modeling of social networking service data on occupational accidents in korea: Latent dirichlet allocation analysis. **J Med Internet Res**, v. 22, n. 8, p. e19222, Aug 2020. ISSN 1438-8871. Available in: <http://www.jmir.org/2020/8/e19222/>.

MäNTYLä, M. V.; GRAZIOTIN, D.; KUUTILA, M. The evolution of sentiment analysisa review of research topics, venues, and top cited papers. **Computer Science Review**, v. 27, p. 16 – 32, 2018. ISSN 1574-0137. Available in: <http://www.sciencedirect.com/science/article/pii/S1574013717300606>.

MOKHTARI, K. E.; CEVIK, M.; BAŞAR, A. Using topic modelling to improve prediction of financial report commentary classes. In: SPRINGER. **Canadian Conference on Artificial Intelligence**. 2020. p. 201–207.

MOREIRA, J.; CESAR, M. R. d. A. Ideologia de Gênero: uma metodologia de análise. **Educação Realidade**, scielo, v. 44, 00 2019. ISSN 2175-6236. Available in: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2175-62362019000400610&nrm=iso>.

MORTENSON, M. J.; DOHERTY, N. F.; ROBINSON, S. Operational research from taylorism to terabytes: A research agenda for the analytics age. **European Journal of Operational Research**, v. 241, n. 3, p. 583 – 595, 2015. ISSN 0377-2217. Available in: <http://www.sciencedirect.com/science/article/pii/S037722171400664X>.

MUKHERJEE, S.; LAMBA, H.; WEIKUM, G. Experience-aware item recommendation in evolving review communities. In: **2015 IEEE International Conference on Data Mining**. 2015. p. 925–930. ISSN 1550-4786.

NASSIRTOUSSI, A. K.; AGHABOZORGI, S.; WAH, T. Y.; NGO, D. C. L. Text mining for market prediction: A systematic review. **Expert Systems with Applications**, v. 41, n. 16, p. 7653 – 7670, 2014. ISSN 0957-4174. Available in: <http://www.sciencedirect.com/science/article/pii/S0957417414003455>.

NAZAR, N.; HU, Y.; JIANG, H. Summarizing software artifacts: A literature review. **Journal of Computer Science and Technology**, v. 31, n. 5, p. 883–909, Sep 2016. ISSN 1860-4749. Available in: <https://doi.org/10.1007/s11390-016-1671-1>.

NIE, W.; LI, X.; LIU, A.; SU, Y. 3d object retrieval based on spatial+lda model. **Multimedia Tools Appl.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 76, n. 3, p. 4091–4104, fev. 2017. ISSN 1380-7501. Available in: <https://doi.org/10.1007/s11042-015-2840-x>.

PAPACHRISTOPOULOS, L.; KLEIDIS, N.; SFAKAKIS, M.; TSAKONAS, G.; PAPATHEODOROU, C. Discovering the topical evolution of the digital library evaluation community. In: GAROUFALLOU, E.; HARTLEY, R. J.; GAITANOU, P. (Ed.). **Metadata and Semantics Research**. Cham: Springer International Publishing, 2015. p. 101–112. ISBN 978-3-319-24129-6.

PERINA, A.; LOVATO, P.; MURINO, V.; BICEGO, M. Biologically-aware latent dirichlet allocation (balda) for the classification of expression microarray. In: DIJKSTRA, T. M. H.; TSIVTSIVADZE, E.; MARCHIORI, E.; HESKES, T. (Ed.). **Pattern Recognition in Bioinformatics**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 230–241. ISBN 978-3-642-16001-1.

POURNARAKIS, D. E.; SOTIROPOULOS, D. N.; GIAGLIS, G. M. A computational model for mining consumer perceptions in social media. **Decision Support Systems**, Elsevier, v. 93, p. 98–110, 2017.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What you need to know about data mining and data-analytic thinking**. : " O'Reilly Media, Inc.", 2013.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Available in: <http://www.R-project.org/>.

ROQUE, C.; CARDOSO, J. L.; CONNELL, T.; SCHERMERS, G.; WEBER, R. Topic analysis of road safety inspections using latent dirichlet allocation: A case study of roadside safety in irish main roads. **Accident Analysis & Prevention**, Elsevier, v. 131, p. 336–349, 2019.

ROSEN-ZVI, M.; GRIFFITHS, T.; STEYVERS, M.; SMYTH, P. The author-topic model for authors and documents. In: AUAI PRESS. **Proceedings of the 20th conference on Uncertainty in artificial intelligence**. 2004. p. 487–494.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 2. ed. : Pearson Education, 2003. ISBN 0137903952.

SANTOS, F. F. d. **Extração de tópicos baseado em agrupamento de regras de associação**. Tese (Doutorado) — Universidade de São Paulo, 2015.

SARASWAT, M.; CHAKRAVERTY, S.; MAHAJAN, N.; TOKAS, N. On using reviews and comments for cross domain recommendations and decision making. In: **2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)**. 2016. p. 3656–3659.

SAURA, J. R.; PALOS-SANCHEZ, P.; GRILO, A. Detecting indicators for startup business success: Sentiment analysis using text data mining. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 11, n. 3, p. 917, 2019.

SBALCHIERO, S.; EDER, M. Topic modeling, long texts and the best number of topics. some problems and solutions. **Quality & Quantity**, Springer, p. 1–14, 2020.

SCHMIEDEL, T.; MüLLER, O.; BROCKE, J. vom. Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. **Organizational Research Methods**, v. 22, n. 4, p. 941–968, 2019. Available in: <https://doi.org/10.1177/1094428118773858>.

SEKIYA, T.; MATSUDA, Y.; YAMAGUCHI, K. Analysis of computer science related curriculum on lda and isomap. In: **Proceedings of the Fifteenth Annual Conference on Innovation and Technology in Computer Science Education**. New York, NY, USA: ACM, 2010. (ITiCSE '10), p. 48–52. ISBN 978-1-60558-820-9. Available in: <http://doi.acm.org/10.1145/1822090.1822106>.

SHANG, L.; CHAN, K.-P. A temporal latent topic model for facial expression recognition. In: KIMMEL, R.; KLETTE, R.; SUGIMOTO, A. (Ed.). **Computer Vision – ACCV 2010**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 51–63. ISBN 978-3-642-19282-1.

SIVIC, J.; RUSSELL, B. C.; EFROS, A. A.; ZISSERMAN, A.; FREEMAN, W. T. Discovering objects and their location in images. In: IEEE. **Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on**. 2005. v. 1, p. 370–377.

SOMMERIA-KLEIN, G.; ZINGER, L.; COISSAC, E.; IRIBAR, A.; SCHIMANN, H.; TABERLET, P.; CHAVE, J. Latent dirichlet allocation reveals spatial and taxonomic structure in a dna-based census of soil biodiversity from a tropical forest. **Molecular Ecology Resources**, Wiley Online Library, v. 20, n. 2, p. 371–386, 2020.

SONG, B.; JIANG, Z.; LIU, L. Automated experiential engineering knowledge acquisition through q&a contextualization and transformation. **Advanced Engineering Informatics**, v. 30, n. 3, p. 467 – 480, 2016. ISSN 1474-0346. Available in: <http://www.sciencedirect.com/science/article/pii/S1474034616301562>.

SRINIVASAN, R.; SENTHILRAJA, M.; INIYAN, S. Pattern recognition of twitter users using semantic topic modelling. In: **2017 International Conference on IoT and Application (ICIOT)**. 2017. p. 1–4.

STA, J.; ZLACKý, D.; HLáDEK, D. Semantically similar document retrieval framework for language model speaker adaptation. In: **2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA)**. 2016. p. 403–407.

STEIGER, E.; WESTERHOLT, R.; RESCH, B.; ZIPF, A. Twitter as an indicator for whereabouts of people? correlating twitter with uk census data. **Computers, Environment and Urban Systems**, v. 54, p. 255 – 265, 2015. ISSN 0198-9715. Available in: <http://www.sciencedirect.com/science/article/pii/S0198971515300181>.

SUKHIJA, N.; TATINENI, M.; BROWN, N.; MOER, M. V.; RODRIGUEZ, P.; CALLICOTT, S. Topic modeling and visualization for big data in social sciences. In: IEEE. **Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences**. 2016. p. 1198–1205.

SUN, C.-Y.; LEE, A. J. Tour recommendations by mining photo sharing social media. **Decision Support Systems**, v. 101, p. 28 – 39, 2017. ISSN 0167-9236. Available in: <http://www.sciencedirect.com/science/article/pii/S0167923617300982>.

SUTHERLAND, I.; SIM, Y.; LEE, S. K.; BYUN, J.; KIATKAWSIN, K. Topic modeling of online accommodation reviews via latent dirichlet allocation. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 12, n. 5, p. 1821, 2020.

TROUSSAS, C.; KROUSKA, A.; VIRVOU, M. Automatic predictions using lda for learning through social networking services. In: **2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)**. 2017. p. 747–751.

TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. **Journal of artificial intelligence research**, v. 37, p. 141–188, 2010.

WANG, X.; GERBER, M. S.; BROWN, D. E. Automatic crime prediction using events extracted from twitter posts. In: **Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction**. Berlin, Heidelberg: Springer-Verlag, 2012. (SBP'12), p. 231–238. ISBN 978-3-642-29046-6.

WANG, X.; MOHANTY, N.; MCCALLUM, A. Group and topic discovery from relations and text. In: ACM. **Proceedings of the 3rd international workshop on Link discovery**. 2005. p. 28–35.

WANG, Y.; LIU, J.; QU, J.; HUANG, Y.; CHEN, J.; FENG, X. Hashtag graph based topic model for tweet mining. In: **2014 IEEE International Conference on Data Mining**. 2014. p. 1025–1030. ISSN 1550-4786.

WENMING, G.; LIHONG, L.; TIANLANG, D. Topic mining for call centers based on a-lda and distributed computing. **Concurrency and Computation: Practice and Experience**, v. 29, n. 3, p. e3776, 2016. E3776 CPE-15-0479.R1. Available in: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3776>.

WHELAN, E.; TEIGLAND, R.; VAAST, E.; BUTLER, B. Expanding the horizons of digital social networks: Mixing big trace datasets with qualitative approaches. **Information and Organization**, v. 26, n. 1, p. 1 – 12, 2016. ISSN 1471-7727. Available in: <http://www.sciencedirect.com/science/article/pii/S1471772716300495>.

WICKHAM, H.; CHANG, W.; WICKHAM, M. H. Package ggplot2. r package version 3.3.2. **Create Elegant Data Visualisations Using the Grammar of Graphics. Version**, v. 2, n. 1, p. 1–189, 2020.

WICKHAM, H.; FRANCOIS, R.; HENRY, L.; MÜLLER, K. Package 'dplyr: A grammar of data manipulation'. r package version 1.0.2. **R Found. Stat. Comput., Vienna. https://CRAN. R-project. org/package= dplyr**, 2020.

WILSON, J.; CHAUDHURY, S.; LALL, B. Clustering short temporal behaviour sequences for customer segmentation using lda. **Expert Systems**, v. 35, n. 3, p. e12250, 2018. E12250 EXSY-Mar-17-055.R1. Available in: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12250>.

XIE, L.; LIU, Y.; CHEN, G. A forensic analysis solution of the email network based on email contents. In: **2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)**. 2015. p. 1613–1619.

YAN, X.; GUO, J.; LIU, S.; CHENG, X.; WANG, Y. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: SIAM. **Proceedings of the 2013 SIAM International Conference on Data Mining**. 2013. p. 749–757.

YANG, P.; WANG, H.; FANG, H.; CAI, D. Opinions matter: a general approach to user profile modeling for contextual suggestion. **Information Retrieval Journal**, v. 18, n. 6, p. 586–610, Dec 2015. ISSN 1573-7659. Available in: <https://doi.org/10.1007/s10791-015-9278-7>.

YE, X.; LI, S.; YANG, X.; QIN, C. Use of social media for the detection and analysis of infectious diseases in china. **ISPRS International Journal of Geo-Information**, v. 5, n. 9, 2016. ISSN 2220-9964.

YEGANOVA, L.; KIM, S.; BALASANOV, G.; WILBUR, W. J. Discovering themes in biomedical literature using a projection-based algorithm. **BMC Bioinformatics**, v. 19, n. 1, p. 269, Jul 2018. ISSN 1471-2105. Available in: <https://doi.org/10.1186/s12859-018-2240-0>.

ZENG, J.; LENG, B.; XIONG, Z. 3-d object retrieval using topic model. **Multimedia Tools and Applications**, v. 74, n. 18, p. 7859–7881, Sep 2015. ISSN 1573-7721. Available in: <https://doi.org/10.1007/s11042-014-2029-8>.

ZGUROVSKY, M. Z.; ZAYCHENKO, Y. P. **Big Data: Conceptual Analysis and Applications**. : Springer, 2020.

ZHANG, C.; ZHU, G.; HUANG, Q.; TIAN, Q. Image classification by search with explicitly and implicitly semantic representations. **Information Sciences**, v. 376, p. 125 – 135, 2017. ISSN 0020-0255. Available in: <http://www.sciencedirect.com/science/article/pii/S0020025516312336>.

ZHANG, H.; CAI, Y.; ZHU, B.; ZHENG, C.; YANG, K.; WONG, R. C.-W.; LI, Q. Incorporating concept information into term weighting schemes for topic models. In: SPRINGER. **International Conference on Database Systems for Advanced Applications**. 2020. p. 227–244.

ZHANG, H.; ZHANG, X.; TIAN, Z.; LI, Z.; YU, J.; LI, F. Incorporating temporal dynamics into lda for one-class collaborative filtering. **Knowledge-Based Systems**, v. 150, p. 49 – 56, 2018. ISSN 0950-7051. Available in: <http://www.sciencedirect.com/science/article/pii/S0950705118300996>.

ZHANG, J. Listening to the consumer: Exploring review topics on airbnb and their impact on listing performance. **Journal of Marketing Theory and Practice**, Taylor & Francis, v. 27, n. 4, p. 371–389, 2019.

ZHANG, Y.; MA, J.; WANG, Z. Semi supervised classification of scientific and technical literature based on semi supervised hierarchical description of improved latent dirichlet allocation (lda). **Cluster Computing**, Jan 2018. ISSN 1573-7543. Available in: <https://doi.org/10.1007/s10586-017-1674-x>.

ZHAO, S.; ZHANG, D.; DUAN, Z.; CHEN, J.; ZHANG, Y.-p.; TANG, J. A novel classification method for paper-reviewer recommendation. **Scientometrics**, v. 115, n. 3, p. 1293–1313, Jun 2018. ISSN 1588-2861. Available in: <https://doi.org/10.1007/s11192-018-2726-6>.

ZHAO, W.; CHEN, J. J.; PERKINS, R.; LIU, Z.; GE, W.; DING, Y.; ZOU, W. A heuristic approach to determine an appropriate number of topics in topic modeling. In: SPRINGER. **BMC bioinformatics**. 2015. v. 16, n. 13, p. S8.

ZHENG, X.; DING, W.; XU, J.; CHEN, D. Personalized recommendation based on review topics. **Service Oriented Computing and Applications**, v. 8, n. 1, p. 15–31, Mar 2014. ISSN 1863-2394. Available in: <https://doi.org/10.1007/s11761-013-0140-8>.

A Appendix 1

Packages used

tm - (FEINERER et al., 2019)

SnowballC - (BOUCHET-VALAT, 2020)

topicmodels - (GRUN et al., 2020)

ggplot2 - (WICKHAM et al., 2020a)

dplyr - (WICKHAM et al., 2020b)

gridExtra - (AUGUIE et al., 2017)

```
0
1  ##########################
2  #      used packages      #
3  ##########################
4
5  library(tm)
6  library(SnowballC)
7  library(topicmodels)
8  library(ggplot2)
9  library(dplyr)
10 library(gridExtra)
11
12
13
14 ##########################
15 #      data processing    #
16 ##########################
17
18 filenames <- list.files(getwd(),pattern = "*.txt")
19 files<-lapply(filenames,readLines)
20 docs<-Corpus(VectorSource(files))
21 writeLines(as.character(docs[[1]]))
22 docs<-tm_map(docs,content_transformer(tolower))
23 toSpace <- content_transformer(function(x,pattern) {return(gsub(pattern,"",x))})
24 docs <- tm_map(docs, removePunctuation)
25 docs <- tm_map(docs, removeNumbers)
26 docs <- tm_map(docs, removeWords, stopwords("english"))
27 docs <- tm_map(docs, stripWhitespace)
28 docs <- tm_map(docs,stemDocument)
29 dtm <- DocumentTermMatrix(docs)
30
31
32
33 ##########################
34 #          metrics       #
35 ##########################
36
37 # Define parameters for Gibbs sampling
38 burnin <- 500
39 iter <- 1000
40 thin <- 500
41 seed <-list(2003,5,63,100001,765)
```

```
nstart <- 5
best <- TRUE


TopicsNumber <- function(dtm, topics = seq(10, 40, by = 10),
                                metrics = "Griffiths2004",
                                method = "Gibbs", control = list(),
                                mc.cores = NA, verbose = FALSE,
                                libpath = NULL) {
# check parameters
  if (length(topics[topics < 2]) != 0) {
    if (verbose) cat("warning: topics count can't to be less than 2, incorrect values was
        removed.\n")
    topics <- topics[topics >= 2]
  }
  topics <- sort(topics, decreasing = TRUE)

  if ("Griffiths2004" %in% metrics) {
    if (method == "VEM") {
      # memory allocation error
      if (verbose) cat("'Griffiths2004' is incompatible with 'VEM' method, excluded.\n")
      metrics <- setdiff(metrics, "Griffiths2004")
    } else {
      # save log-likelihood
      if (!"keep" %in% names(control)) control <- c(control, keep = 50)
    }
  }

# fit models
  if (verbose) cat("fit models...")

# Parallel setup
  if (any(class(mc.cores) == "cluster")) {
    cl <- mc.cores
  } else if (isTRUE(class(mc.cores) == "integer")) {
    cl <- parallel::makeCluster(mc.cores)
  } else {
    cl <- parallel::makeCluster(parallel::detectCores())
  }
  parallel::setDefaultCluster(cl)
  parallel::clusterExport(varlist = c("dtm", "method", "control"),
                          envir = environment())
  models <- parallel::parLapply(X = topics, fun = function(x) {
    if (is.null(libpath) == FALSE) { .libPaths(libpath) }
    topicmodels::LDA(dtm, k = x, method = method, control = control)
  })
  if (! any(class(mc.cores) == "cluster")) {
    parallel::stopCluster(cl)
  }
  if (verbose) cat(" done.\n")

# calculate metrics
  if (verbose) cat("calculate metrics:\n")
  result <- data.frame(topics)
  for(m in metrics) {
    if (verbose) cat(sprintf("  %s...", m))
    if (! m %in% c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014")) {
      cat(" unknown!\n")
    } else {
      result[m] <- switch(m,
```

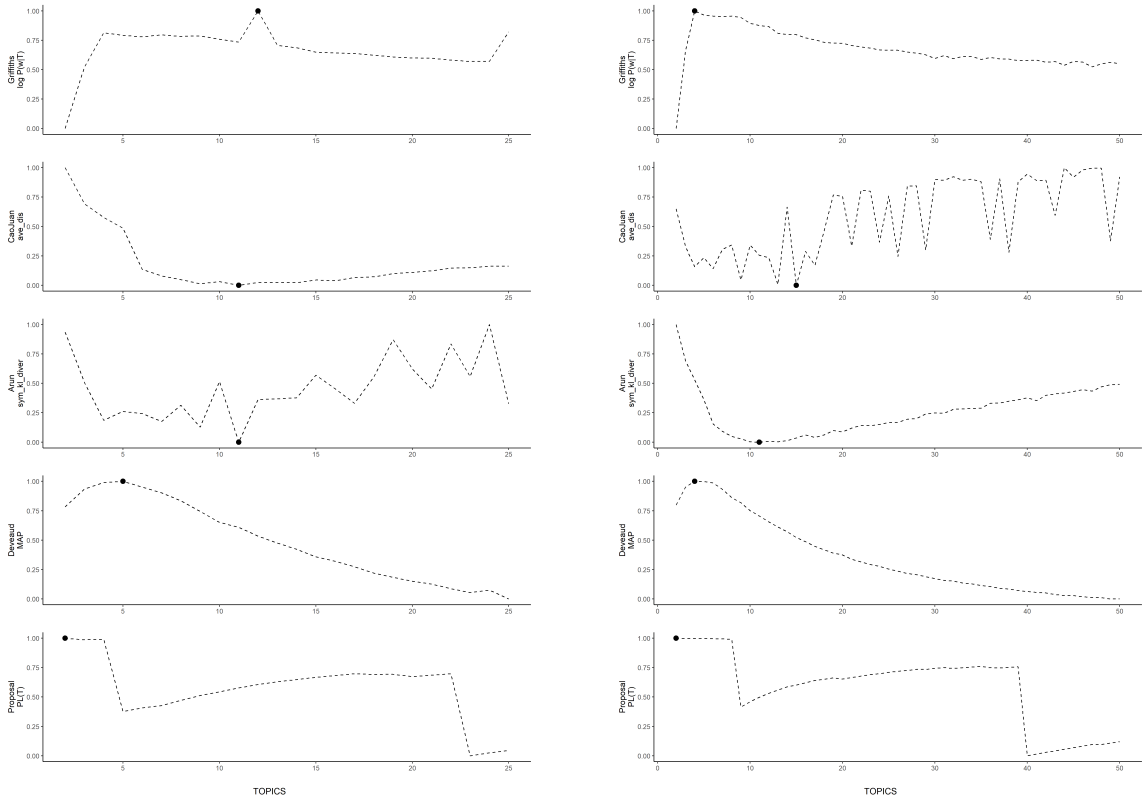
```
                            "Griffiths2004" = Griffiths2004(models, control),
                            "CaoJuan2009"   = CaoJuan2009(models),
                            "Arun2010"      = Arun2010(models, dtm),
                            "Deveaud2014"   = Deveaud2014(models),
                            NaN
        )
        if (verbose) cat(" done.\n")
      }
    }

    return(result)
}


ptm <- proc.time()

#' @keywords internal
Griffiths2004 <- function(models, control) {
  # log-likelihoods (remove first burning stage)
  burnin  <- ifelse("burnin" %in% names(control), control$burnin, 0)
  logLiks <- lapply(models, function(model) {
    utils::tail(model@logLiks, n = length(model@logLiks) - burnin/control$keep)
    # model@logLiks[-(1 : (control$burnin/control$keep))]
  })
  # harmonic means for every model
  metrics <- sapply(logLiks, function(x) {
    # code is a little tricky, see explanation in [Ponweiser2012 p. 36]
    # ToDo: add variant without "Rmpfr"
    llMed <- stats::median(x)
    metric <- as.double(
      llMed - log( Rmpfr::mean( exp( -Rmpfr::mpfr(x, prec=2000L) + llMed )))
    )
    return(metric)
  })
  return(metrics)
}
proc.time() - ptm


ptm <- proc.time()
CaoJuan2009 <- function(models) {
  metrics <- sapply(models, function(model) {
    # topic-word matrix
    m1 <- exp(model@beta)
    # pair-wise cosine distance
    pairs <- utils::combn(nrow(m1), 2)
    cos.dist <- apply(pairs, 2, function(pair) {
      x <- m1[pair[1], ]
      y <- m1[pair[2], ]
      # dist <- lsa::cosine(x, y)
      dist <- crossprod(x, y) / sqrt(crossprod(x) * crossprod(y))
      return(dist)
    })
    # metric
    metric <- sum(cos.dist) / (model@k*(model@k-1)/2)
    return(metric)
  })
  return(metrics)
}
proc.time() - ptm
```

```
161
162 ptm <- proc.time()
163 Arun2010 <- function(models, dtm) {
164   # length of documents (count of words)
165   len <- slam::row_sums(dtm)
166   # evaluate metrics
167   metrics <- sapply(models, FUN = function(model) {
168     # matrix M1 topic-word
169     m1 <- exp(model@beta) # rowSums(m1) == 1
170     m1.svd <- svd(m1)
171     cm1 <- as.matrix(m1.svd$d)
172     # matrix M2 document-topic
173     m2  <- model@gamma    # rowSums(m2) == 1
174     cm2  <- len %*% m2    # crossprod(len, m2)
175     norm <- norm(as.matrix(len), type="m")
176     cm2  <- as.vector(cm2 / norm)
177     # symmetric Kullback-Leibler divergence
178     divergence <- sum(cm1*log(cm1/cm2)) + sum(cm2*log(cm2/cm1))
179     return ( divergence )
180   })
181   return(metrics)
182 }
183 proc.time() - ptm
184
185
186 ptm <- proc.time()
187 Deveaud2014 <- function(models) {
188   metrics <- sapply(models, function(model) {
189     ### original version
190     # topic-word matrix
191     m1 <- exp(model@beta)
192     # prevent NaN
193     if (any(m1 == 0)) { m1 <- m1 + .Machine$double.xmin }
194     # pair-wise Jensen-Shannon divergence
195     pairs  <- utils::combn(nrow(m1), 2)
196     jsd <- apply(pairs, 2, function(pair) {
197       x <- m1[pair[1], ]
198       y <- m1[pair[2], ]
199       ### standard Jensen-Shannon divergence
200       # m <- (x + y) / 2
201       # jsd <- 0.5 * sum(x*log(x/m)) + 0.5 * sum(y*log(y/m))
202       ### divergence by Deveaud2014
203       jsd <- 0.5 * sum(x*log(x/y)) + 0.5 * sum(y*log(y/x))
204       return(jsd)
205     })
206
207     #     ### optimized version
208     #     m1   <- model@beta
209     #     m1.e <- exp(model@beta)
210     #     pairs  <- utils::combn(nrow(m1), 2)
211     #     jsd <- apply(pairs, 2, function(pair) {
212     #       x   <- m1[pair[1], ]
213     #       y   <- m1[pair[2], ]
214     #       x.e <- m1.e[pair[1], ]
215     #       y.e <- m1.e[pair[2], ]
216     #       jsd <- ( sum(x.e*(x-y)) + sum(y.e*(y-x)) ) / 2
217     #       return(jsd)
218     #     })
219
220     # metric
```

```
    metric <- sum(jsd) / (model@k*(model@k-1))
    return(metric)
  })
  return(metrics)
}
proc.time() - ptm

############################
#          Proposal        #
############################


#number of topics
k<-100

ptm <- proc.time()
#lda
ldaOut <-LDA(dtm,k, method="Gibbs", control=list(nstart=nstart, seed = seed, best=best,
    burnin = burnin, iter = iter))
posterior<-posterior(ldaOut, control=list(nstart=nstart, seed = seed, best=best, burnin =
     burnin, iter = iter))
write.csv(posterior[["terms"]],"posterior.csv")
DADOSDF<-t(posterior[["terms"]])
write.csv(DADOSDF,"DADOSDF.csv")


DADOS_X1<-DADOSDF[,c(1:1)]
DADOS_X2<-DADOSDF[,c(1:2)]
DADOS_X3<-DADOSDF[,c(1:3)]
DADOS_X4<-DADOSDF[,c(1:4)]
...
DADOS_X98<-DADOSDF[,c(1:98)]
DADOS_X99<-DADOSDF[,c(1:99)]
DADOS_X100<-DADOSDF[,c(1:100)]

DADOS_X1<-as.numeric(DADOS_X1)
DADOS_X2<-as.numeric(DADOS_X2)
DADOS_X3<-as.numeric(DADOS_X3)
...
DADOS_X98<-as.numeric(DADOS_X98)
DADOS_X99<-as.numeric(DADOS_X99)
DADOS_X100<-as.numeric(DADOS_X100)


X1<-harmonic.mean(DADOS_X1)
X2<-harmonic.mean(DADOS_X2)
X3<-harmonic.mean(DADOS_X3)
...
X98<-harmonic.mean(DADOS_X98)
X99<-harmonic.mean(DADOS_X99)
X100<-harmonic.mean(DADOS_X100)


X1<-X1$harmean
X2<-X2$harmean
X3<-X3$harmean
...
X98<-X98$harmean
X99<-X99$harmean
X100<-X100$harmean
```

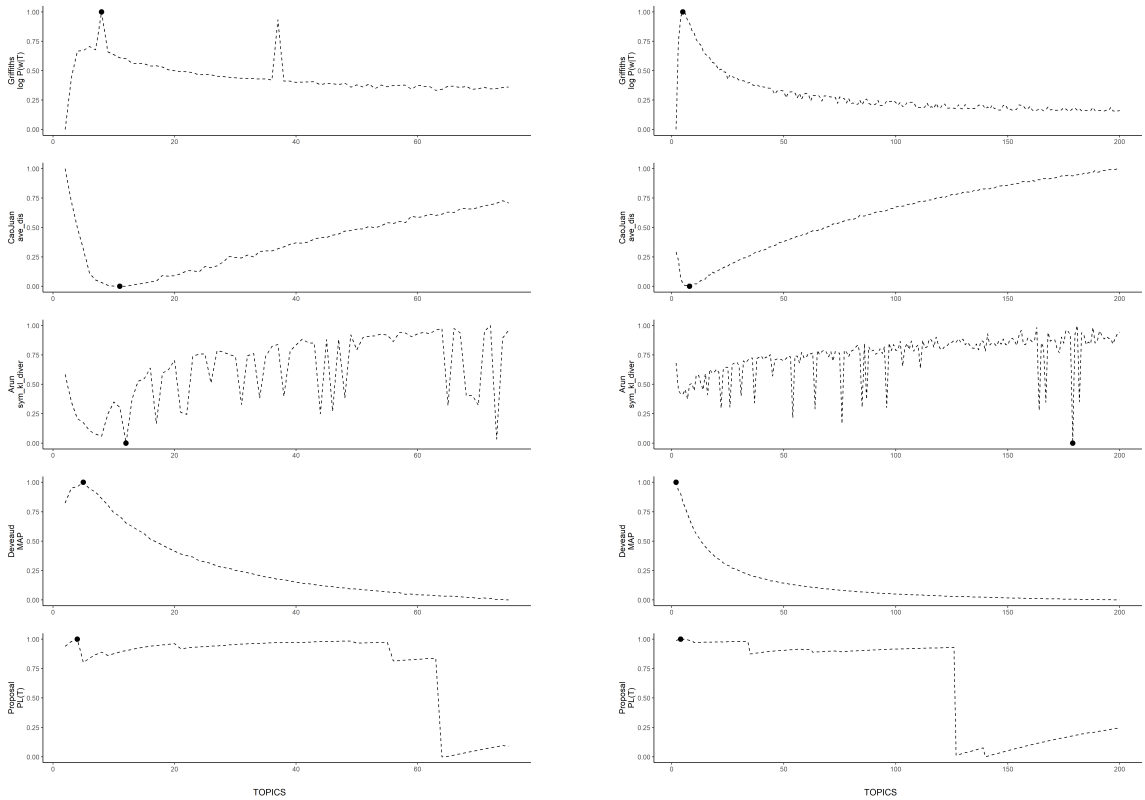
```
279
280 X1<-log(X1)
281 X2<-log(X2)
282 X3<-log(X3)
283 ...
284 X98<-log(X98)
285 X99<-log(X99)
286 X100<-log(X100)
287 media_harmonica<-
288
289 data.frame(topico=c(1,2,3,4,5,...,99,100),
290 mediaharmonica=c(X1,X2,X3,X4,X5,...,X99,X100))
291 media_harmonica<-media_harmonica[order(media_harmonica$mediaharmonica,decreasing=T),]
292 proc.time() - ptm
293
294 View(media_harmonica)
295 write.csv(media_harmonica,"media_harmonica.csv")
```

B Appendix 2



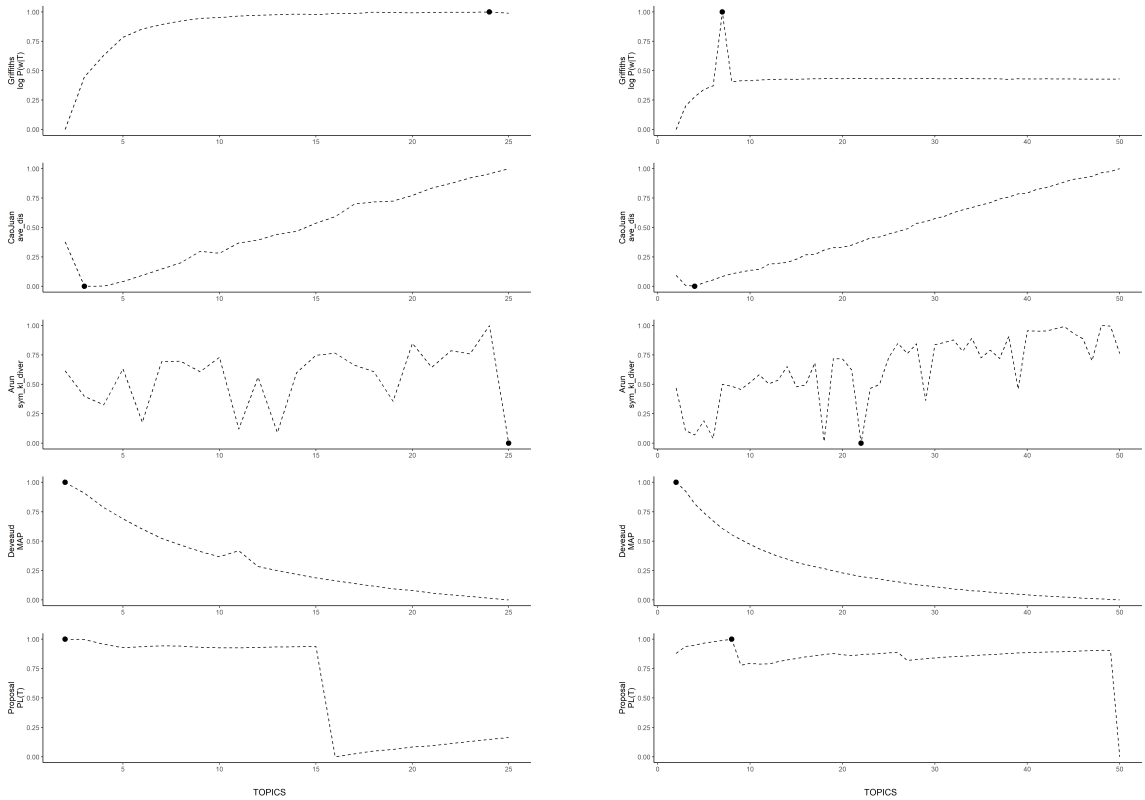
(a) input K of 2 to 25 (b) input K of 2 to 50

Figure 18 – Experimental results of the behavior of the different metrics - NYT articles dataset



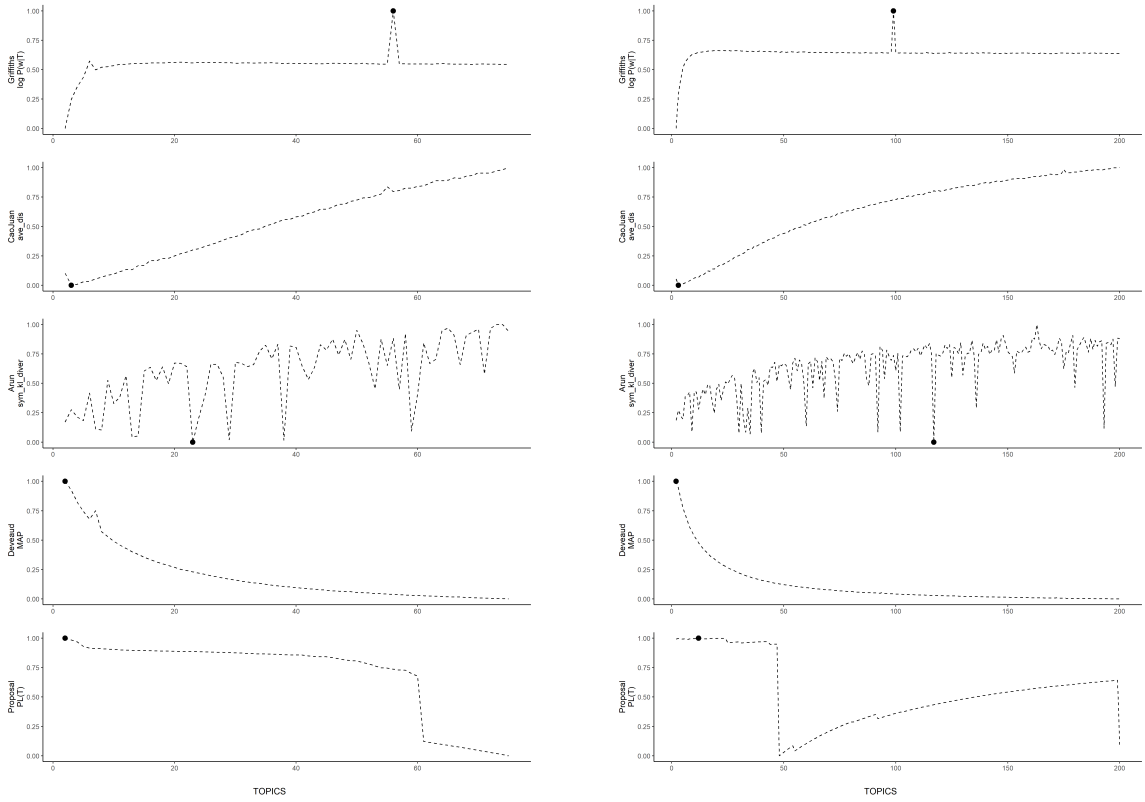
(a) input K of 2 to 75 (b) input K of 2 to 200

Figure 19 – Experimental results of the behavior of the different metrics - NYT articles dataset



(a) input K of 2 to 25 (b) input K of 2 to 50

Figure 20 – Experimental results of the behavior of the different metrics - Customer reviews dataset



(a) input K of 2 to 75 (b) input K of 2 to 200

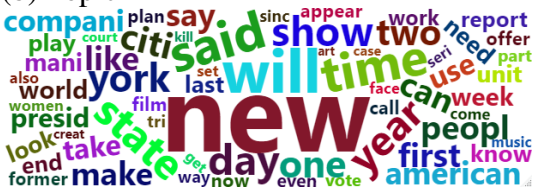
Figure 21 – Experimental results of the behavior of the different metrics - Customer reviews dataset

C Appendix 3

(a) Topic 1

(b) Topic 2

(c) Topic 3



(d) Topic 4

(e) Topic 5

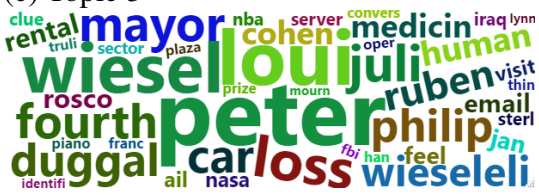
Figure 22 – Word clouds with the 100 most relevant words in each topic - NYT articles dataset - **Deveaud** [2 to 100].

(a) Topic 1

(b) Topic 2

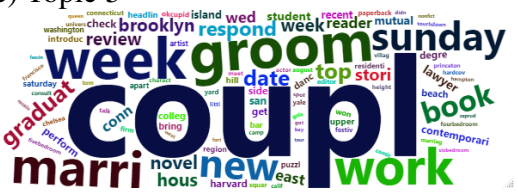
(c) Topic 3

(d) Topic 4

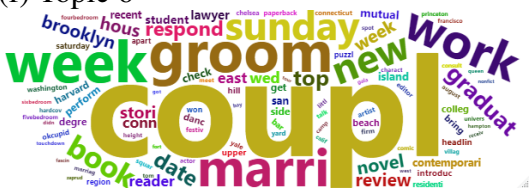


(e) Topic 5

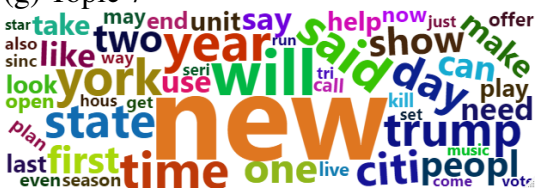
(f) Topic 6



(g) Topic 7



(h) Topic 8



(i) Topic 9

Figure 23 – Word clouds with the 100 most relevant words in each topic - NYT articles dataset - **Griffths** [2 to 100].

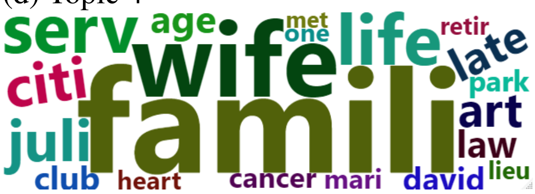
(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

(e) Topic 5



(f) Topic 6

(g) Topic 7

(h) Topic 8

(i) Topic 9

(j) Topic 10

(k) Topic 11

(l) Topic 12

Figure 24 – Word clouds with the 100 most relevant words in each topic - NYT articles dataset - **CaoJuan** [2 to 100].

(a) Topic 1

(b) Topic 2

(c) Topic 3

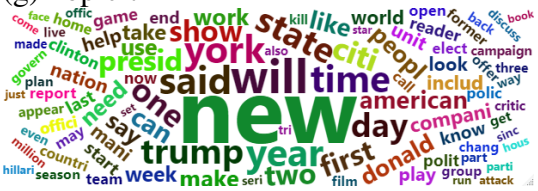
(d) Topic 4

(e) Topic 5

(f) Topic 6

(g) Topic 7

(h) Topic 8



(i) Topic 9

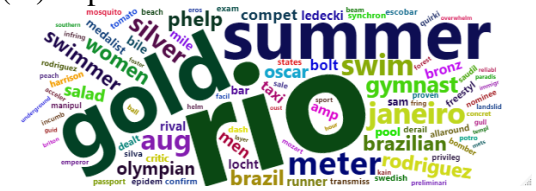
(j) Topic 10

(k) Topic 11

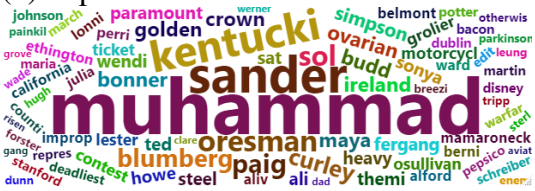
(l) Topic 12

(m) Topic 13

(n) Topic 14



(o) Topic 15



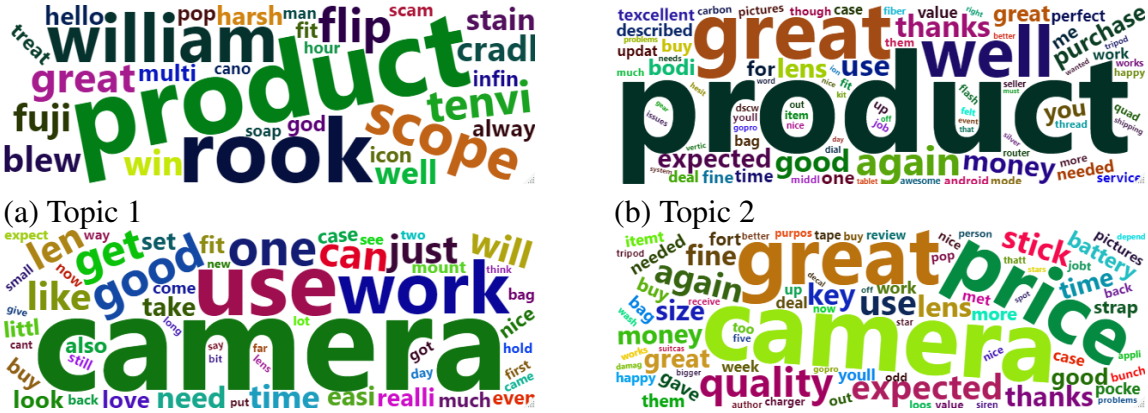
(p) Topic 16

Figure 25 – Word clouds with the 100 most relevant words in each topic - NYT articles dataset - **Arun** [2 to 100].

D Appendix 4

(a) Topic 1 (b) Topic 2

Figure 26 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Deveaud** [2 to 100].



(a) Topic 1

(b) Topic 2

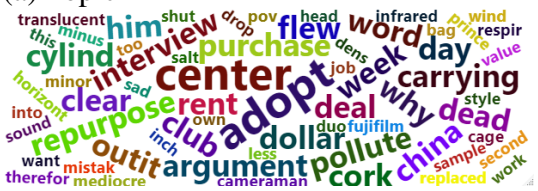
(c) Topic 3

(d) Topic 4

Figure 27 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **CaoJuan** [2 to 100].

(a) Topic 1

(b) Topic 2



(c) Topic 3

(d) Topic 4

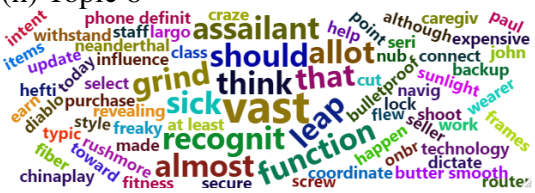
(e) Topic 5

(f) Topic 6

(g) Topic 7

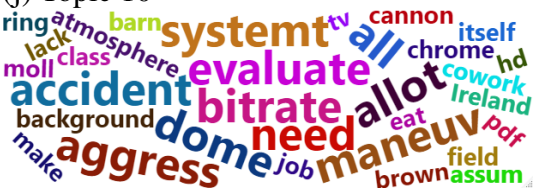
(h) Topic 8

(i) Topic 9



(j) Topic 10

(k) Topic 11



(l) Topic 12

(m) Topic 13

(n) Topic 14

(o) Topic 15

(p) Topic 16

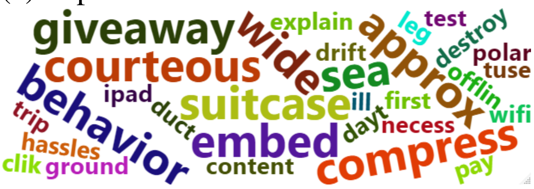
Figure 28 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Griffths** [2 to 100] - Topics 1 to 16 of 52.

(a) Topic 17

(b) Topic 18

(c) Topic 19

(d) Topic 20



(e) Topic 21

(f) Topic 22

(g) Topic 23

(h) Topic 24

(i) Topic 25

(j) Topic 26

(k) Topic 27

(l) Topic 28

(m) Topic 29

(n) Topic 30

(o) Topic 31

(p) Topic 32

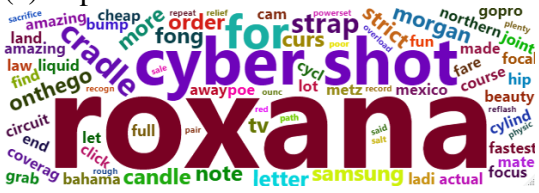
Figure 29 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Griffths** [2 to 100] - Topics 17 to 32 of 52.

(a) Topic 33

(b) Topic 34

(c) Topic 35

(d) Topic 36



(e) Topic 37

(f) Topic 38

(g) Topic 39

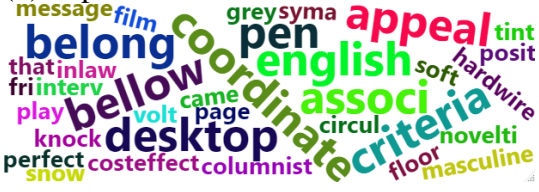
(h) Topic 40

(i) Topic 41

(j) Topic 42

(k) Topic 43

(l) Topic 44



(m) Topic 45

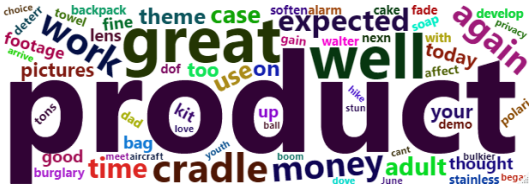
(n) Topic 46

(o) Topic 47

(p) Topic 48

Figure 30 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Griffths** [2 to 100] - Topics 33 to 48 of 52.

(a) Topic 49



(b) Topic 50

(c) Topic 51

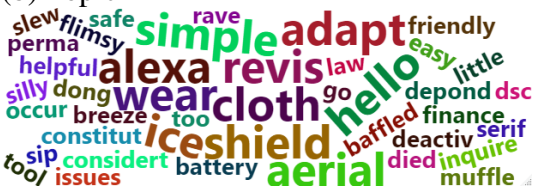
(d) Topic 52

Figure 31 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Griffths** [2 to 100] - Topics 49 to 52 of 52.

(a) Topic 1

(b) Topic 2

(c) Topic 3

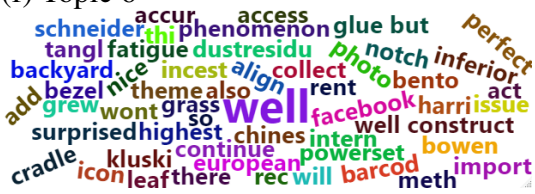


(d) Topic 4

(e) Topic 5

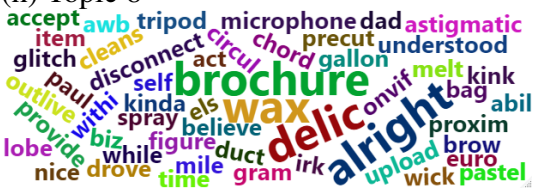
(f) Topic 6

(g) Topic 7



(h) Topic 8

(i) Topic 9



(j) Topic 10

(k) Topic 11

(l) Topic 12

(m) Topic 13

(n) Topic 14

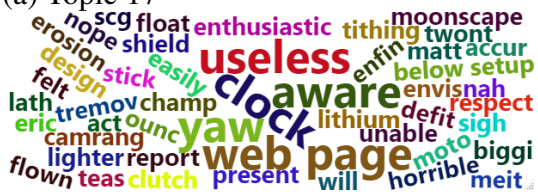
(o) Topic 15

(p) Topic 16

Figure 32 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Arun** [2 to 100] - Topics 1 to 16 of 78.

(a) Topic 17

(b) Topic 18



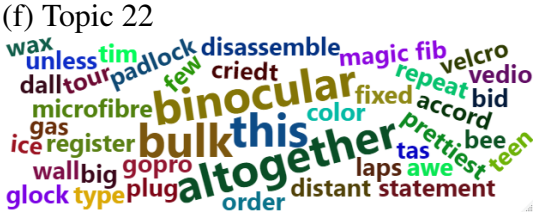
(c) Topic 19

(d) Topic 20

(e) Topic 21

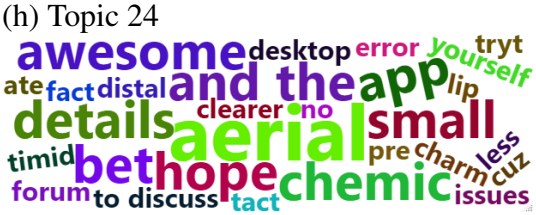
(f) Topic 22

(g) Topic 23



(h) Topic 24

(i) Topic 25



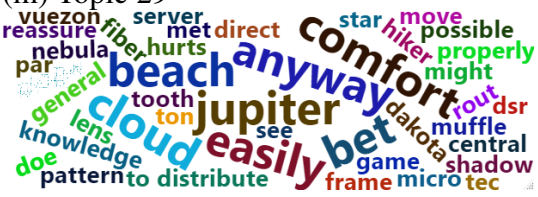
(j) Topic 26

(k) Topic 27

(l) Topic 28

(m) Topic 29

(n) Topic 30



(o) Topic 31

(p) Topic 32

Figure 33 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Arun** [2 to 100] - Topics 17 to 32 of 78.

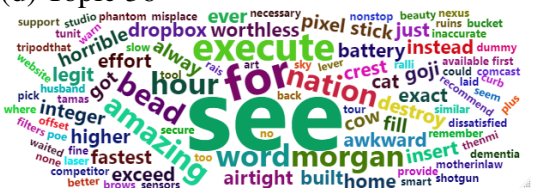
(a) Topic 33

(b) Topic 34

(c) Topic 35

(d) Topic 36

(e) Topic 37



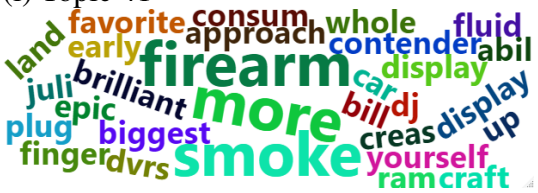
(f) Topic 38

(g) Topic 39

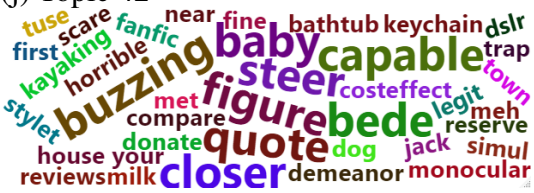
(h) Topic 40

(i) Topic 41

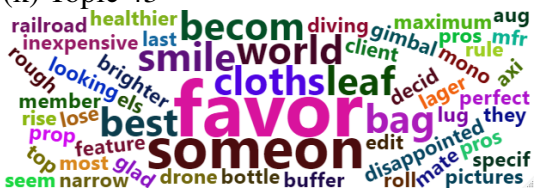
(j) Topic 42



(k) Topic 43

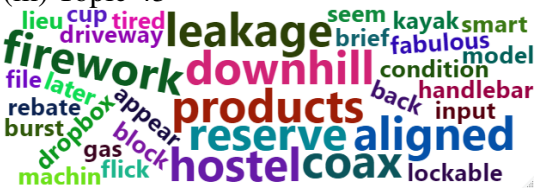


(l) Topic 44



(m) Topic 45

(n) Topic 46



(o) Topic 47

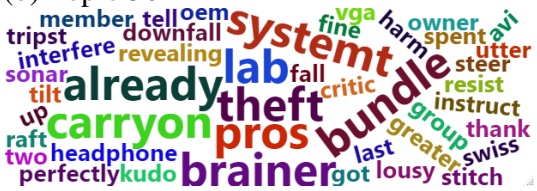
(p) Topic 48

Figure 34 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Arun** [2 to 100] - Topics 33 to 48 of 78.

(a) Topic 49

(b) Topic 50

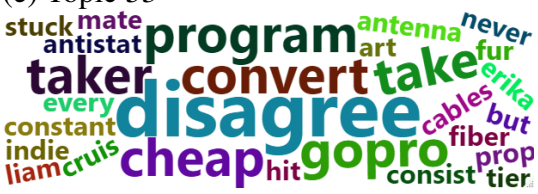
(c) Topic 51



(d) Topic 52

(e) Topic 53

(f) Topic 54



(g) Topic 55

(h) Topic 56

(i) Topic 57

(j) Topic 58

(k) Topic 59

(l) Topic 60

(m) Topic 61

(n) Topic 62

(o) Topic 63

(p) Topic 64

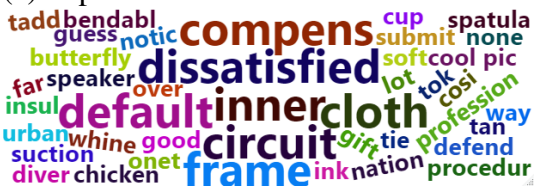
Figure 35 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Arun** [2 to 100] - Topics 49 to 64 of 78.

(a) Topic 65

(b) Topic 66

(c) Topic 67

(d) Topic 68



(e) Topic 69

(f) Topic 70

(g) Topic 71

(h) Topic 72

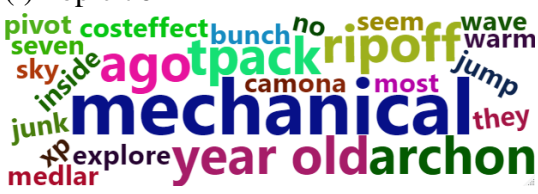
(i) Topic 73

(j) Topic 74

(k) Topic 75

(l) Topic 76

(m) Topic 77



(n) Topic 78

Figure 36 – Word clouds with the 100 most relevant words in each topic - Amazon customer reviews dataset - **Arun** [2 to 100] - Topics 65 to 78 of 78.