Alternative for selecting the number of topics in Latent Dirichlet Allocation

Abstract - A topic model is based on a corpus of documents, discovers the topics that permeate the corpus and assigns documents to those topics. In this article, we propose a metric to select the number of topics used in the Latent Dirichlet Allocation (LDA) model. 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.

Keywords: Latent Dirichlet Allocation, Topic model, Categorical data

1 INTRODUCTION

With the continued growth of various information sources, where a large amount of data is being generated every minute, the need for data compression, analysis tools and management is becoming evident. According to Davenport (2014), in 2012, the world used over 2.8 trillion gigabytes of data. Companies that could take advantage of this large volume of data, converting them into ideas, innovations, and commercial value, are only scratching the surface of what is possible. The same study points out that only 0.50% of the 2.8 trillion gigabytes is analyzed. The development of hardware and software platforms, mainly for the web network, has allowed the rapid creation of large amounts and different types of data, particularly text data.

Based on Chen, Chiang and Storey (2012), computational tools are required to organize, search, and understand large amounts of data, especially as our collective knowledge continues to be digitized and stored by many types of media. Regarding textual data, Srivastava, Salakhutdinov and Hinton (2013) highlight that 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, discovering patterns in word usage, and to connect documents that share similar patterns. Pattern discovery reflects the underlying topics that come together to form the documents (Han, Pei & Kamber, 2014).

The representation of the original source plays a key role in defining topics and identifying which ones are present in each document This results in a clear representation of documents that is useful for analyzing the subjects present within them (Miner, 2012). Topic extraction models can be divided into non-probabilistic and probabilistic.

Probabilistic models have gained popularity following the introduction of the probabilistic latent semantic analysis (pLSA) model by Hofmann (1999), the first to formalize the extraction of probabilistic topics. Although it provides a good basis for the of texts, the pLSA model presents two problems. First, the topic generation process for each document, which requires determining a quantity of parameters that grows linearly with the number of documents and can lead to overfitting of the estimated parameters, is not defined. In addition, the pLSA model does not define a natural way to calculate probabilities related to a document not in the training set (Aggarwal & Zhai, 2012; Blei, NG & Jorndan, 2003; Kim, Park, Lu & Zhai, 2012).To avoid these problems, Blei et al. (2003) proposed the Latent Dirichlet Allocation (LDA) model, which we will focus on in this paper. It is the most popular probabilistic topic model.

The LDA model is conditioned by three parameters: the Dirichlet hyperparameters (α and β) and the number of topics (K). According to Cao, Xia, Li, Zhang and Tang (2009) points 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. According to Arun, Suresh, Madhavan and Murthy (2010) shows that the challenge is ensuring that a small number of latent topics are sufficient to effectively represent a large corpus. Almost all topic modeling methods assume implicitly that the number of topics is known in advance. The determination of the parameter K is extremely important and little explored in the literature, mainly due to the intensive and delayed computation procedure.

In this sense, the main goal of this paper is to develop a metric to identify the ideal number for the parameter K of the LDA model, which allows for an adequate representation of the corpus and an agile computation procedure.

The remaining sections of this paper are organized as follows. In Section 2, we review the basic principles of LDA and some proposed methods for choosing the number of topics in LDA. In Section 3, we propose our approach and show the experimental results in Section 4. Finally, conclusions were drawn and a proposal for future work is presented in Section 5.

2 RELATED WORK

2.1 Latent Dirichlet Allocation (LDA)

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 hyperparameters, 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 into different topicswhich will fill in the documents. One can guess the meaning of the name latent Dirichlet allocation, which expresses the intention of the model, to allocate the latent topics, obeying the distribution of Dirichlet (Arun et al., 2010; Blei, 2012; Cao et al., 2009).

The intuitive idea behind the LDA, illustrated by the authors of Blei et al. (2003), is the assumption that several topics, distributions of words, exist for the entire collection. Each document is assumed to be generated as follows: First, distribution on the topics is chosen and then, for each word, a topic assignment and the corresponding topic word are chosen. The authors exemplify this idea in their seminal work on the use of data analysis to determine the number of genes an organism needs to survive (in an evolutionary sense). This is done by analyzing the article and different words used in the article, such as computer, forecast, life, organism, genes, sequenced. 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 in different proportions. LDA is a statistical model of document collections that attempts to capture this intuitive grouping (Blei, 2012).

The LDA generation 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. Using LDA, 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 (Blei et al., 2003).

The whole generative process can be represented graphically with a Bayesian network. This network is illustrated in the Fig 1.

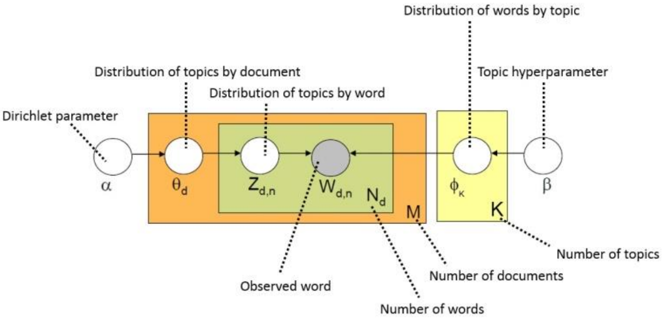

Figure 1. Graphical model of the LDA (Blei et al., 2003).

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.The distribution of the topics for each document are on the second level. The distribution of topics for words within a single document and those repeated from the last level are on the third level. It is possible to represent a document as a mixture of topics (Blei, 2012; Blei et al., 2003).

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 (Arun et al., 2010; Cao et al., 2009; Blei, 2012; Blei et al., 2003).

Then, with the distributions ϕ and θ_d, the document d_j is generated. In the LDA model, a document is considered to simply be a bag of words, with n_d terms in a document d. The terms in the bag of words are vocabulary words, and occasionally, repetitions of the same word may occur. For each position i of the n_d term positions of a bag of words, a word from 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 (Blei et al., 2003).

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} stores the chosen topic. Then, the ϕ distribution is chosen as the word that fills position i. The variable ϕ is K n-dimensional distributions, where each distribution K, ϕ_K, corresponds to proportions of words that semantically describe the subject that topic K treats. The term w_{dn} should be chosen from the topic z_{dn}, obeying the word distribution $\phi_{z_{dn}}$ (Arun et al., 2010; Blei et al., 2003; Cao et al., 2009).

2.2 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 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. The main constituent of this latter probability is the probability of the data provided in the model, integrating all parameters in the model. In other words, it requires the training of several LDA models to select the one with the best performance. It is an intensive and time-consuming computation procedure. In Griffiths and Steyvers (2004), they propose the production of a set of models by changing the K parameter. The focus is on analyzing $P(w|K)$, where w is a word in the corpus. According to Griffiths and Steyvers (2004) points out that the challenge is to obtain the sum of all possible word assignments to topics z. They solve this problem using an approximation of $P(w|K)$, obtaining the harmonic mean of a set of values of $P(w|z, K)$.

According to Griffiths and Steyvers (2004) describes the behavior of a 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.

According to 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. 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 $Density\ (Z, r)$.

The average cosine distance of the model $r1 = ave\_dis(\beta)$, 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 $Cardinality$: $K_{n+1} = K_n + f(r) \times (K_n - C_n) \cdot r$ is guided by $f(r)$. If it is negative $f_{n+1}(r) = -1 \times f_n(r)$, else $f_{n+1}(r) = -f_n(r) \cdot f_0(r) = -1$ (Cao et al., 2009).

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 (Cao et al., 2009).

According to 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) are the results of 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. 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.

According to Deveaud, SanJuan and Bellot (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 metric proposed by Deveaud et al. (2014) always seeks the highest values.

There are other approaches in the literature to determine the K parameter in topic models, such as Albuquerque, Valle and Li (2019), Cheng, He and Liu (2015) and Taddy (2012). This study will not be considered.

3 THE METHOD PROPOSAL 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, $\mathrm{P}(\mathrm{w}|k)$, $\mathrm{k} = 1,2,\ldots,\mathrm{K}$. We have that $\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{P}(\mathrm{w}|\mathrm{k}) = 1$ 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

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

where nw is the number of words in the corpus. $P(w|T_n)$ 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 $\mathrm{K} = n^*$.

4 EXPERIMENTS

To better understand the proposal, we will 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. Experiments were run on a 12 GB RAM computer.

4.1 First Dataset

The first dataset used was composed of news from The New York Times, an American newspaper based in New York City with worldwide influence and readership. This dataset contains 42,139 news articles with 41,162 different words. We computed 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.

Our first experiment was carried out on the entire corpus described above, and the results are shown in Table 1.

Table 1
Experimental results with different metrics and different K ranges – First Dataset.

METRICS	TOPICS				
	2 to 25	*2 to 50*	*2 to 75*	*2 to 100*	*2 to200*
Griffths	12	4	8	9	5
CaoJuan	11	11	11	12	8
Arun	11	15	12	16	179
Deveaud	5	4	5	5	2
Proposal	2	2	4	6	4

We observed that 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 observethat at the higher ranges of the interval, the values of the best K of Griffths and Proposal are similar.

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

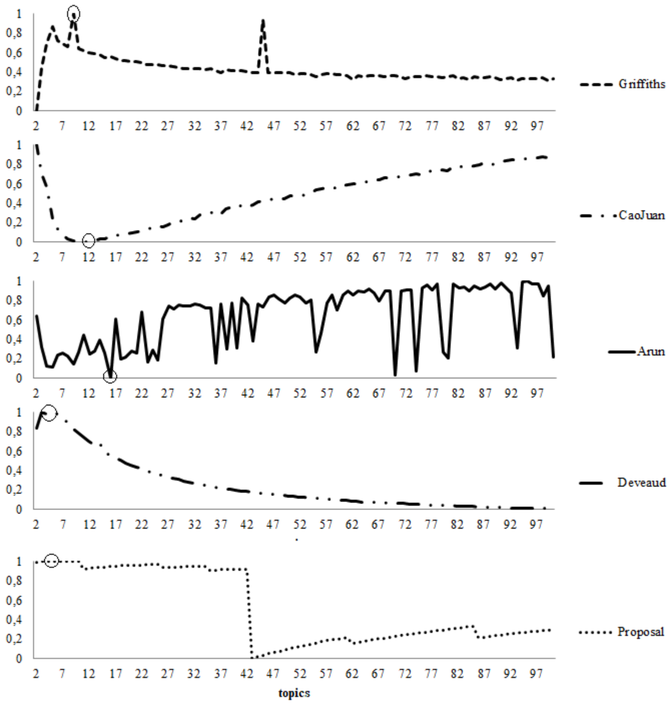

Figure 2. Experimental results of the behavior of the different metrics with an input of K = 100 – First Dataset.

In Figure 3, we show the processing time of the experiment. It is evident that the best performance is by the Proposal metric. The performance of 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.

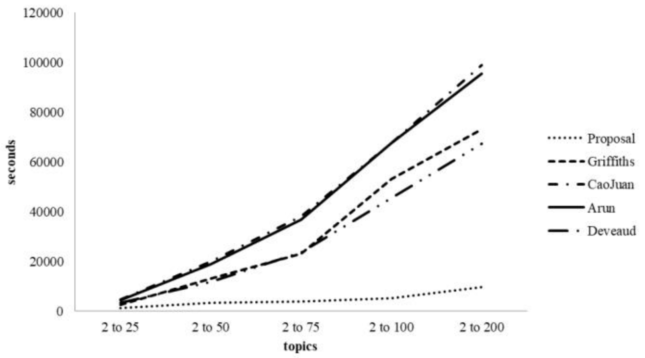

Figure 3. Processing time – First Dataset.

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.

4.2 Second Dataset

The first dataset used was composed of Amazon customer reviews (camera product), contains 25,000 Customer Reviews with 20,510 different words. We computed the 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 and 2 to 100.

The result of the second experiment that was carried out with the entire corpus described above is shown in Table 2.

Table 2
Experimental results with different metrics and different K ranges – Second Dataset.

METRICS	**TOPICS**			
	2 to 25	***2 to 50***	***2 to 75***	***2 to 100***
Griffths	24	7	56	41
CaoJuan	3	4	3	4
Arun	25	22	23	78
Deveaud	2	2	2	2
Proposal	2	8	6	9

The [Arun] metric again presents a significant discrepancy when the range's amplitude increases, as the metric considers word, topic and document information the range's amplitude considered influences the construction of the matrices and affects the results. [Deveaud] remains unchanged in the different intervals, its objective is to find the most dispersed model, which indicates that when considering the model with only 2 topics, this characteristic is found.

Just as in the first experiment the [CaoJuan] metric remains stable and how he considers the correlation between topics is indicative that there is little difference in correlation when the number of K increases. In [Griffiths], a behavior very similar to the first experiment is observed, where there is an oscillation when changing the input range, as the metric indicates that there is this characteristic behavior indicating that with smaller intervals the metric does not reach the so-called peak. The Proposal metric presents a stable behavior suggesting an intermediate number of K considering the suggestion of the other metrics.

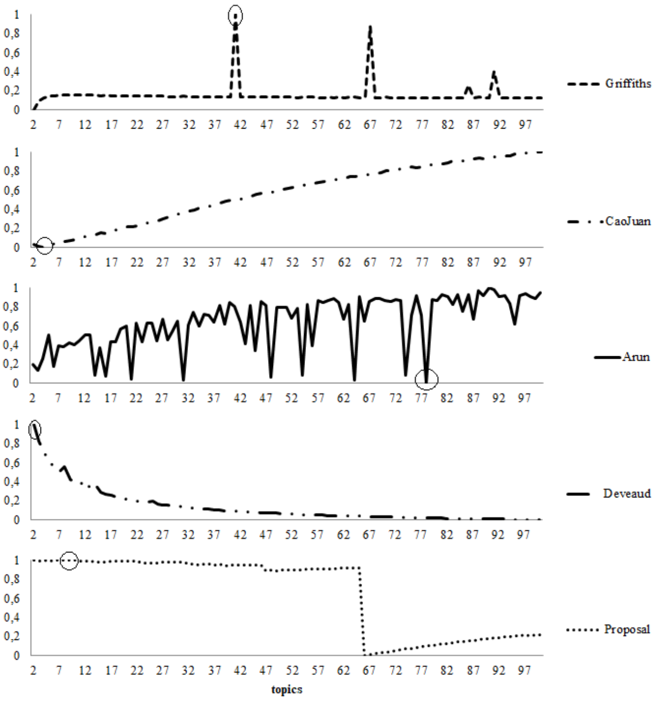

Figure 4. Experimental results of the behavior of the different metrics and with K=100 input – Second Dataset.

In Fig 4 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.

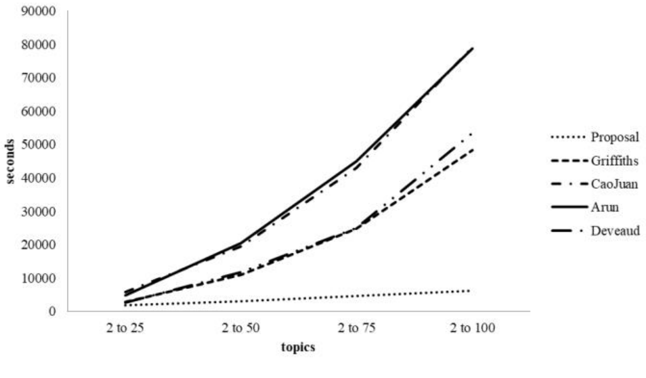

Figure 5. Processing time – Second Dataset.

The processing time is somewhat limiting, because in the [Griffiths] metric there is an indication that it is necessary to increase the K amplitude to perform more tests, as well as in [Arun], where there is no certainty of the behavior, especially when analyzing the graphs (Fig 2 and Fig 4). In Fig 5 it is evident that the processing time of the Proposal metric is better than the other metrics

5 CONCLUSIONS

In the LDA model, the number of topics is a parameter fundamental to the construction of the model, as it is reflected in the analysis of the data. The objective of this article was to propose a metric with which to determine the number of topics in the LDA model and to compare its performance with that of other metrics, mainly taking the processing time into consideration.

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

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 that this method is effective. Future explorations include tests on other topic models and image data.

Acknowledgment

The authors gratefully acknowledge the sponsorship of Coordination for the Improvement of Higher Education Personnel (CAPES).

References

Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media.

Albuquerque, P. H., Valle, D. R., & Li, D. (2019). Bayesian LDA for mixed-membership clustering analysis: The Rlda package. *Knowledge-Based Systems*, 163, 988-995.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9), 1775-1781.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities.* Harvard Business Review Press.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1), 61-84.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). *Business intelligence and analytics: From big data to big impact.* MIS quarterly, 1165-1188.

Cheng, D., He, X., & Liu, Y. (2015, February). Model selection for topic models via spectral decomposition. In *Artificial Intelligence and Statistics* (pp. 183-191).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228-5235.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57).

Kim, H. D., Park, D. H., Lu, Y., & Zhai, C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology*, 49(1), 1-10.

Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Srivastava, N., Salakhutdinov, R. R., & Hinton, G. E. (2013). Modeling documents with deep boltzmann machines. *arXiv preprint arXiv:1309.6865.*

Taddy, M. (2012, March). On estimation and selection for topic models. In *Artificial Intelligence and Statistics* (pp. 1184-1193).