Journal of Statistical and Econometric Methods, Vol. 11, No. 1, 2022, 1-14
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1111
Scientific Press International Limited

Determining the Number of Topics to Retain using Tools from Factor Analysis

Holmes Finch[1]

Abstract

Determining the optimal number of topics to retain in the conduct of topic modeling (TM) has received much attention over the last decade. Despite this work, issues remain regarding the best methods to use for making such determinations. Approaches involving the use of relatively simple statistics, most notably perplexity, have proven to be somewhat inconsistent. Recently, researchers have suggested the use of change in perplexity scores as a useful heuristic for determining the optimal number of topics to retain. The current study builds on this earlier work by assessing the utility of several methods borrowed from factor analysis and applied to statistics commonly used in topic modeling, including perplexity and Alpha. These new approaches are applied to several textual datasets and compared with more traditional methods for determining the number of topics to retain. Results of these analyses demonstrate that application of these methods borrowed from factor analysis does appear to be effective for identifying the number of topics to retain.

[1] Department of Educational Psychology, Ball State University, Muncie, USA.

Article Info: *Received:* February 25, 2020. *Revised:* June 23, 2020.
Published online: January 6, 2022.

1. Introduction

Topic modeling (TM) is widely used by researchers in many fields to identify a relatively small number of topics underlying a collection of documents, based on the pattern of word co-occurrence in a corpus of texts. The resulting topics can provide researchers with insights into themes in a body of text, as well as with how words are typically used to communicate these themes. Examples of TM applications can be found in fields as varied as literature (Jockers & Mimno, 2013), history (Thompson, Batista-Navarro, Kontonatsios, Carter, Toon, McNaught, Timmerman, Worboys, & Ananiadou, 2016), political science (Ficcadenti, Cerqueti, & Ausloos, 2019), medicine (Piedra, Ferrer, & Gea, 2014), and business (Klevak, Livnat, & Suslava, 2019), among others. TMs are typically fit using a Latent Dirichlet Allocation (LDA) models, as described in Blei, Ng, and Jordan (2003). This model, which will be discussed in more detail below, relies on the Markov Chain Monte Carlo (MCMC) estimator in order to obtain parameter estimates.

A key issue in the conduct of TM using LDA is determination of the number of topics to retain. A variety of approaches have been suggested for this purpose, but research remains very much open regarding the optimal approach. The purpose of this study was to build upon prior work in this area through the application of methods for determining the number of latent variables to retain in actor analysis to the problem of identifying the optimal number of topics for a given corpus of textual data. In the following sections, traditional methods for determining the optimal number of topics are described, followed by a discussion of an alternative paradigm based upon statistics commonly used in exploratory factor analysis (EFA) to determine the number of latent traits that are present. Next, the specifics of how these methods are applied to LDA and topic modeling will be described. Finally, these methods, as well as more traditional approaches are applied to several sets of textual data and results are compared to one another, and to the standard methods.

2. Methods

2.1 Latent Dirichlet Allocation

One of the most common approaches for fitting topic models to a corpus of texts is LDA (Blei, Ng, & Jordan, 2003). Researchers using LDA make an assumption that underlying the observed word counts in a corpus of documents is a finite number of unobserved topics. These topics are associated with specific words, and each document in the corpus is in turn assumed to be associated with the topics to varying degrees, as reflected by the distribution of words in each. In order to identify these latent topics, LDA maximizes the probability of the observed textual data, *D*, given a finite set of topics and parameters associated with the TM as follows:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\sum_{n=1}^{N_d} p\left(z_{dn}|\theta_d p\left(w_{d_n}|z_{dn},\varphi\right)p(\varphi|\beta)\right)\right) d\,\theta_d d\varphi \qquad (1)$$

Where
M =Total number of documents
N_d =Number of words in document *d*
θ_d =Multinomial distribution of topics for document *d* with parameter α
φ =Multinomial distribution of words by topic, with parameter β
z_{dn} =Specific topic taken from θ_d
w_{d_n} =Specific word taken from φ.

Thus, in the context of TM, LDA searches the parameter space of α and β in order to maximize the probability of the observed documents and their associated word distributions, *D*.
Exact estimation of the model in equation (1) is intractable, and thus the Markov Chain Monte Carlo (MCMC) estimator is used to obtain posterior distributions for the parameter of interest. Prior distributions for α and β must be provided, and the resultant model will yield posterior distributions for each of these. These posterior distributions provide researchers with the primary estimates of interest, namely the association of specific words with topics, and the distribution of topics across documents. The Gibbs sampler can be used in the conduct of MCMC, which was the case in this study.

2.2 Using perplexity to determine the number of topics to retain

There exist a number of statistics that can be used by researchers to determine the number of topics to retain when using TM. One of the more common of these is perplexity, which is defined for number of topics *j* as

$$P_j = exp\left\{-\frac{\sum_{d=1}^{M} ln(p(w_d))}{\sum_{d=1}^{M} N_d}\right\} \qquad (2)$$

Where
M =Number of documents
N_d =Number of words in document *d*
$p(w_d)$ =Probability of word *w* in document *d*.

Researchers using perplexity typically fit several models with differing numbers of topics to a test set of documents from the corpus. The model is then applied to a cross-validation set of documents from the corpus, and the TM solution with the lowest perplexity value is selected as optimal (Zhao, et al., 2015). Although relatively simple to apply, this approach has been shown to be relatively unstable with respect to identifying replicable TM solutions (REFERENCE).
An alternative approach for determining the number of topics to retain was suggested by Zhao, et al., (2015) for determining the number of topics to retain

involves the rate of perplexity change (RPC). With this approach, the difference in perplexity values is calculated for each pair of adjacent number of topics (e.g., 2 vs 3, 3 vs 4, etc.). These differences reflect the rate of change in perplexity as the number of topics is increased. The researcher using this approach would then identify the point at which the RPC increases in value, thereby marking the optimal number of topics to retain from a TM analysis. The RPC statistic for *j* topics can be expressed formally as:

$$RPC_j = \left|\frac{P_j - P_{j-1}}{t_j - t_{j-1}}\right| \quad (3)$$

Where
t_j =Number of topics for solution j.

The optimal number of topics is then selected to be that for which $RPC_j < RPC_{j+1}$. Using 3 text datasets drawn from genomics and drug side effects listings, Zhao, et al. compared the RPC approach to the use of the minimum of P_j. Their results demonstrated that the RPC method provided more stable and accurate results with respect to identifying the optimal number of topics than did finding the minimum of P_j. In addition, they argued that the use of RPC_j offers researchers with a simpler approach than was the case when using the minimum P_j criterion as a part of conducting a full sensitivity analysis.

2.3 Using Alpha to determine the number of topics to retain

In addition to perplexity, it is also possible to use the document topic density, as measured by the Alpha statistic, to determine the number of factors to retain. A greater mixture of topics within documents (i.e., documents contain more topics) is associated with larger values of Alpha, whereas when documents are primarily associated with only 1 topic, Alpha approaches 0. Thus, when conducting TM, researchers may wish to find a solution for which the value of Alpha is minimized, indicating that each document in the corpus is likely to be associated with a small number of topics.

2.4 Factor analysis based methods

An alternative paradigm for determining the number of topics using P_j, RPC_j, and Alpha is based on a set of approaches that were developed for use in EFA. These methods are designed to help researchers use the eigenvalues from an EFA to determine the number of factors to retain. One such approach is the optimal coordinate (OC) test (Raiche, et al., 2012), which compares the actual eigenvalue for a given factor (e.g., factor 3) with the eigenvalue that would be predicted using a two-point regression model based on the set of eigenvalues obtained from the covariance matrix of the observed data (the optimal coordinate). The two points used in the regression equation for predicting eigenvalue *i* would be the $(i+1)^{th}$

eigenvalue, and the last eigenvalue. If the observed eigenvalue *i* is larger than the predicted eigenvalue associated with factor *i*, then factor *i* is retained. The researcher would examine each of these comparisons and retain factors up to the first one for which the observed eigenvalue was less than the predicted eigenvalue. In the context of TM, rather than eigenvalues, P_j, RPC_j, or Alpha would be included in the two point regression analysis.

A second eigenvalue based test from EFA is the acceleration factor (AF) test. The acceleration factor statistic is calculated as the second derivative of the regression equation used to predict the optimal coordinate, as described above. This second derivative is then applied to each eigenvalue in order to calculate the acceleration factor, which is simply a measure of the steepness in the line connecting the points in the scree plot. The last factor to be retained is the one that precedes the coordinate where the acceleration factor is maximized. As with the optimal coordinates approach, P_j, RPC_j, and Alpha can be substituted for the eigenvalues from EFA.

In addition to the optimal coordinate and acceleration factor methods, there are two other objective approaches based on eigenvalues that have been discussed in the EFA literature, and which have proven to be effective in simulation studies. Gorsuch's (1983) CNG scree test involves the calculation of the slope linking the first three eigenvalues, then the calculation of the slope linking eigenvalues 2, 3, and 4, then the slope linking eigenvalues 3, 4, and 5, and so on. The researcher then compares these slopes with one another, and selects the number of factors where the difference between the slopes is greatest. Thus, for example, if the largest difference between slope values lies between the line for points 2, 3, and 4 versus the line for points 3, 4, and 5, we would retain 4 factors.

Zoski and Jurs (1993) suggested a variant of the Gorsuch approach, known as NMREG, in which pairs of regression equations are estimated using all of the data points, rather than just sets of 3 at a time. Thus, for *p* indicator variables, the following pairs of equations would be considered:

Line 1 (eigenvalues 1, 2, and 3)
Line 2 (eigenvalues 4 through *p*)
Line 3 (eigenvalues 1, 2, 3, and 4)
Line 4 (eigenvalues 5 through *p*)
Line 5 (eigenvalues 1, 2, 3, 4, and 5)
Line 6 (eigenvalues 6 through *p*)

The slopes for the lines in each pair (e.g., line 1 versus line 2) are then compared using a *t*-test, and the number of factors to be retained is associated with the maximum *t* value. As an example, if the maximum *t* statistic is associated with the comparison between lines 3 and 4, then 4 factors (corresponding the largest factor number in line 3) would be retained. As noted above, simulation research has shown that NMREG, and the CNG test are both very effective at determining the number of factors to retain, assuming that there are at least 3 latent variables present in the

data (Raiche, et al., 2012). And as with the optimal coordinates and acceleration factor methods, eigenvalues can be replaced by P_j, RPC_j, and Alpha.

2.5 Using measures of topic distance to determine the number of topics to retain

In addition to using P_j, RPC_j, or Alpha for determining the number of topics to retain, the literature also features the use of other alternatives for the purpose of identifying the optimal number of topics to retain. One family of such approaches, known collectively as LDA tuning, is based upon the calculation of distance between pairs of topics in a corpus (Arun, et al., 2010; Cao, et al., 2009). Thus for a given number of topics, an LDA model is fit to the data and the composite distance among all topics is calculated, based on dissimilarity in the sharing of terms. In other words, topics with more differences in terms of the probability of being associated with specific words will exhibit greater distance from one another. The optimal number of topics is that which maximizes the distances among topics. Differences in the methods representing this approach are associated with how distance is calculated. The interested reader is referred to Arun, et al. and Cao, et al. for these details.

Deveaud, et al., (2014) described an alternative approach for determining the number of topics to retain based on the LDA model. They define the optimal number of topics as that which satisfies the following equation:

$$\widehat{K} = argmax\left(\frac{1}{K(K-1)}\right)\sum_{(k,k')\in T_k} D(k \parallel k') \quad (4)$$

Where
K =Number of topics
T_k =The full set of topics modeled by LDA
$D(k \parallel k')$ =The Jensen-Shannon divergence between topics k and k'.

The Jensen-Shannon divergence is a symmetrized version of the Killback-Leibler distance, and is calculated as:

$$D(k \parallel k') = 0.5\sum_{w\in W_k\cap W_{k'}} P_{TM}(w|k)ln\left(\frac{P_{TM}(w|k)}{P_{TM}(w|k')}\right) +$$

$$0.5\sum_{w\in W_k\cap W_{k'}} P_{TM}(w|k')ln\left(\frac{P_{TM}(w|k')}{P_{TM}(w|k)}\right) \quad (5)$$

Where
$P_{TM}(w|k)$ =Probability of word *w* appearing in topic *k*.

Thus, the optimal number of topics is that which maximizes the distances among the topics, expressed as the probability of specific words appearing in each. Using multiple text corpora, Deveaud, et al. demonstrated that their approach to identifying the optimal number of topics to retain was effective and accurate, particularly when larger numbers of topics were present.

3. Application of methods to textual data

3.1 Description of datasets and analysis

In order to assess and compare the performance of these methods for determining the optimal number of topics to retain, they were applied to three different corpora. Each of these textual datasets was selected so as to have a clearly delineated expected number of topics, based on content. The first dataset included 11 Wikipedia articles on the following topics: Integrals, Riemann integrals, Riemann-Stieltjes integrals, Derivatives, limit of a sequence, Edvard Munch, Vincent Van Gogh, Jan Matejko, Lev Tolstoy, Franz Kafka, and J.R.R. Tolkien. The second TM analysis was conducted on 4 books taken from the online Gutenberg project corpus, including Pride and Prejudice, Twenty Thousand Leagues Under the Sea, Great Expectations, and The War of the Worlds. The final set of data included online threads from 4 USENET discussion groups, including those focused on nutrition, space science, computers, and computer encryption. These three examples were selected for two reasons. First, they represent texts of varying lengths, from novels, to multi-paragraph Wikipedia entries, to short texts discussing particular topics with multiple authors. Second, each includes documents that can be differentiated through the appearance of clearly delineated topics.

All data analyses described below were conducted using the R software package, version 3.6.2 (R Core Team, 2019). TM was done with the TM and Topic models packages, whereas the LDA tuning approaches were carried out using the LDA tuning R library. For each corpus, LDA was carried out, as described in more detail below, and the P_j, RPC_j, and Alpha were calculated. The various heuristics for determining the number of topics to retain were then applied to these statistics, including those based on EFA. The R library nFactors was used for this latter purpose, and applied to P_j, RPC_j, and Alpha.

3.2 Wikipedia results

As noted above, the Wikipedia dataset included a total of 11 documents, covering topics in science, mathematics, and literature. LDA was fit for models of 2 to 20 topics, and the approaches for determining the optimal number of topics to retain described above were used. The minimum $P_j$value was associated with 11 topics, and the RPC_j values increase between 3 and 4 topics, suggesting that 3 topics should be retained, based on the heuristic described by Zhang, et al. (2015). The number of topics to be retained based upon the objective factor analysis based methods, and the LDA tuning approaches appear in Table 1.

Table 1: Number of Topics by Method for Wikipedia Corpus

Statistic	Number of topics
NMREG for perplexity	11
NMREG for perplexity change	5
NMREG for Alpha	11
CNG for perplexity	9
CNG for perplexity change	4
CNG for Alpha	3
OC for perplexity	1
OC for perplexity change	1
AF for perplexity	7
AF for perplexity change	7
Devaud	7
Arun	15
Cao	11

These results cover a wide array of potential topic numbers, from 1 to 11. Recall that the Wikipedia corpus used in this study included 11 separate entries, suggesting a priori the possibility that 11 topics is optimal.

In order to validate the results displayed in Table 1, the topics were characterized by the terms and documents associated with them. In the interest of parsimony, the two most common solutions (7 and 11) are presented here. Table 2 displays the Wikipedia document by the topic for which it had the highest probability based on word content.

Table 2: Wikipedia Document Topic Assignment for 7 and 11 Topics

Document	Topic Assignment: 7 Topics	Topic Assignment: 11 Topics
Integral	3	1
Riemann Integral	3	2
Riemann-Stieltjes Integral	3	3
Derivative	3	4
Limit of a sequence	3	5
Edvard Munch	5	6
Vincent Van Gogh	7	7
Jan Matejko	5	8
Lev Tolstoy	1	9
Franz Kafka	2	10
J.R.R. Tolkien	2	11

When 11 topics were retained, each document was placed into its own topic. On the other hand, when 7 topics were retained, the Wiki documents focused on mathematics were placed into the same topic (3), the articles on Edvard Munch and

Jan Mateiko appeared together (topic 5), as did the articles on Kafka and Tolkien (topic 2). Finally, the Wikis on Van Gogh (7) and Tolstoy (1) were each in a topic of their own. No documents were most associated with topics 4 or 6.
The 5 most common terms associated with each topic appear in Table 3.

Table 3: Most Common Words Associated with each Topic for 7 and 11 Topics

Topic	Most common words: 7 Topics	Most common words: 11 Topics
1	Tolstoy, Russian, Articles, War, Leo	Integral, Derivative, Identifiers, Calculus, Leibnitz
2	Kafka, Tolkien, English, Rings, Lord	Riemann, Function, Interval, Limit, Definition
3	Integral, Function, Derivative, Riemann, Integrations	Generalization, Riemann, Stieltjes, Integral, Function
4	Work, Series, Amsterdam, April, Left	Derivative, Calculus, Graph, Approximation, Differentiation
5	Munch, Museum, Majejko, Portraits, Paintings	Limit, Sequence, Definition, Number, Series
6	Articles, Film, Literary, Modern, View	Munch, Museum, Paintings, Art, Legacy
7	Van Gogh, Vincent, Art, Theo, Museum	Van Gogh, Vincent, Art, Theo, Museum
8		Matejko, Krakow, Polish, Portraits, Museum
9		Tolstoy, Russian, Articles, War, Leo
10		Kafka, Bohemia, Write, Troubled, Czech
11		Tolkien, Rings, Lord, Trilogy, Writer

For the 11 topic solution, the associated words clearly reflect the contents of the articles associated with them. For example, Topic 1 was most associated with the article on Integrals, and likewise the 5 most common terms associated with Topic 1 were Integral, Derivative, Identifiers, Calculus, and Leibnitz, all of which are key concepts associated with integrals. A similar pattern is evident for the other 10 topics associated with this solution. For the 7 topic solution, all of the mathematics focused articles were placed together, and the most commonly associated terms were indeed focused on these topics. With respect to the other topics, however, the patterns of most common words are somewhat less clear. For example, Topic 2 was associated with articles on Kafka and Tolkien, both of whom were writers in the early to mid 20th century. However, outside of their names, the most commonly associated words did not appear to be associated with both individuals. Three of these words were Lord, Rings, and English, all of which were very characteristic of Tolkien, who was English, and who wrote the Lord of the Rings Trilogy, but not of the Czech writer Kafka. In addition, topics 4 and 6 were characterized by terms that may not reflect a coherent theme, at least in comparison to the most common words associated with some of the other topics, or with those in the 11 topics solution.

Given their identification of the presence of 11 topics, and the fact that 11 topics appears to be an optimal solution based upon the grouping of texts and terms, the results presented above provide support for the use of NMREG for P_j and for Alpha, as well as the Cao index. It is not clear that the other heuristics used with this example identified a coherent set of topics.

3.3 Gutenberg Results

The P_j value was minimized for 9 topics, suggesting that this is the optimal number to retain. The RPC_j values for the Gutenberg book topics increased between 2 and 3, indicating that 2 topics should be retained. Table 4 includes the number of topics to be retained based upon the indices included in this study.

Table 4: Number of Topics by Method for Gutenberg Corpus

Statistic	Number of topics
NMREG for perplexity	4
NMREG for perplexity change	4
NMREG for Alpha	4
CNG for perplexity	6
CNG for perplexity change	4
CNG for Alpha	3
OC for perplexity	1
OC for perplexity change	1
AF for perplexity	3
AF for perplexity change	4
Devaud	6
Arun	10
Cao	10

NMREG for perplexity, perplexity change, and Alpha, as well as CNG for perplexity change, and AF for perplexity change all suggest the presence of 4 topics. Given that 4 separate novels were included in this analysis, the conclusion that 4 topics are present appears to be warranted. In order to investigate this issue further, topic membership for each text, as well as the most common words associated with each topic were identified. Book by topic appears in Table 5. Each text is associated with its own topic, supporting the distinct nature of the topics.

Table 5: Novel Topic Assignment for 4 Topics

Document	Topic Assignment: 4 Topics
Great Expectations	1
Pride and Prejudice	2
War of the Worlds	3
Twenty Thousand Leagues under the Sea	4

The words most commonly associated with each topic appear in Table 6.

Table 6: Five Most Common Words Associated with Topics for 4 Topic Solution

Topic	Most common words: 7 Topics
1	Joe, Pip, Havisham, Wemmick, Time
2	Elizabeth, Darcy, Jane, Bennet, Lady
3	Night, People, Martians, Dark, Dead
4	Captain, Nautilus, Sea, Nemo, Water

The most common terms associated with Topic 1 primarily correspond to proper names in the book Great Expectations, which is the text associated with this topic. Likewise, the most common words in Topics 2 and 4 are either proper names, or terms closely associated with the books aligned with each topic. Finally, the most common words linked to Topic 3 include Martians, Dead, and Night, which can be seen as key plot elements in the War of the Worlds. In short, the most common terms associated with each topic are clearly linked to the book belonging to that topic. Taken together, these results appear to support the 4 topic solution, and thus provide validity evidence for the use of NMREG for perplexity, perplexity change, and Alpha, as well as for CNG and AF for perplexity change.

3.4 USENET Results

LDA models were fit to the USENET data for from 2 to 12 topics, with an expectation based on content that 4 topics might be optimal. The minimum perplexity value appeared for 9 topics, whereas the change in RPC_j suggested the presence of 4 topics. Table 7 includes the number of topics identified by each of the statistics included in this study.

Table 7: Number of Topics by Method for USENET Corpus

Statistic	Number of topics
NMREG for perplexity	4
NMREG for perplexity change	4
NMREG for Alpha	4
CNG for perplexity	3
CNG for perplexity change	5
CNG for Alpha	3
OC for perplexity	5
OC for perplexity change	1
AF for perplexity	1
AF for perplexity change	1
Devaud	6
Arun	12
Cao	12

The NMREG for P_j, RPC_j, and Alpha all identified the presence of 4 topics, which is the expected number for the USENET corpus. None of the other methods indicated the presence of 4 topics, although results for the CNG statistics, as well as OC for perplexity suggested the presence of either 3 or 5 topics.
Table 8 includes the most common terms associated with the 4 topics, based upon the LDA modeling.

Table 8: Five Most Common Words Associated with Topics for 4 Topic Solution: USENET

Topic	Most common words: 7 Topics
1	People, DB, water, food, msg
2	Key, chip, encryption, government, bit
3	Data, information, system, software, computer
4	Space, time, NASA, science, power

The first topic appears to be associated primarily with food terms, whereas the second topic is associated with encryption, the third with computers, and the fourth with space. Thus, we can see from these results that the TM has successfully identified the USENET science discussion groups associated with nutrition, encryption, computers, and space. As was the case with the other analyses described above, the NMREG statistic applied to perplexity, perplexity change, and Alpha was able to identify the number of topics expected to be in the data, given the content of the corpus. The other approaches were not as accurate in this case.

4. Conclusions

The goal of this study was to describe several new approaches for ascertaining the optimal number of topics to retain in the context of LDA for TM. These new approaches adapt several statistics that are used to determine the number of factors to retain in EFA. The results of the examples conducted above suggest that the NMREG statistic, when applied to P_j, RPC_j, and Alpha may be a useful tool for accurately identifying the number of topics to retain when using LDA with a corpus of texts. In those example, texts of differing types and lengths were used, and in each case, these three approaches were able to accurately identify the number of topics that corresponded to meaningful topics, both in terms of the most common words, and the organization of the documents themselves. The relative ease of applying NMREG using the R nFactors package, coupled with this accuracy would seem to make it a very useful tool for researchers working with TM.
Future research should continue to vet the methods used here. Though the current set of results does seem to support the use of P_j, RPC_j, and Alpha in conjunction with NMREG, other corpora of texts should be examined. For example, it is unclear how well these models might work with very short texts, or very large corpora. The examples chosen for this study were meant to be representative of those typical

in many areas of research. However, it is certainly true that a wider array of such examples should be pursued. In addition, future research should also examine the use of other statistics common to TM, particularly entropy, in conjunction with the EFA based approaches featured here.

It is hoped that the current work will prove to be useful to researchers working with TM in a wide array of disciplines. The determination of the optimal number of topics to retain has proven to be a challenge, with a number of approaches being suggested for this purpose. The current study provides evidence that combining commonly used statistics from the world of TM, such as P_j, RPC_j, and Alpha, with objective methods designed for use in the context of EFA may give researchers a useful set of tools for determining the number of topics to retain. These methods are easy to implement and interpret, and based upon the work described above, accurate for this purpose.

References

[1] Arun, R., Suresh, V., Veni Madhavan, C.E., & Murthy Narasimha, M.N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In M.J. Zaki, J.X. Yu, B. Ravindran, & V. Pudi, Advances in Knowledge Discovery and Data Mining. London, Springer.

[2] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

[3] Cao, Y., Xia, T., Le, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72, 1775-1781.

[4] Deveaud, R., San Juan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numerique 17, 1, 61-84.

[5] Ficcadenti, V., Cerqueti, R., & Ausloos, M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. Expert Systems with Applications, 123(1), 127-142.

[6] Gorsuch, R.L. (1983). Factor Analysis. Hillsdale, NJ: Erlbaum.

[7] Jockers, M. & Mimno, D. (2013). Significant themes in 19th Century literature. Poetics, 41(6), 750-769.

[8] Klevak, J., Livnat, J., & Suslava, K. (2019). A practical approach to advanced text mining in finance. The Journal of Financial Data Science, 1(1), 122-129.

[9] Pedra, D., Ferrer, A., & Gea, J. (2014). Text mining and medicine: Usefulness in respiratory diseases. Archivos de Bronconeumologia, 50(3), 113-119.

[10] R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[11] Raiche, G., Wall, T.A., Magis, D., Riopel, M., & Blais, J.-G. (2012). Non-graphical solutions for Cattell's scree test. Methodology, 9(1), 23-29.

[12] Thompson, P., Batista-Navarro, R.T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmerman, Ca., Worboys, M., & Ananiadou, S., (2016). Text Mining the history of medicine. PLOS ONE, 11(1).

[13] Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics, 16, 58-67.

[14] Zoski, K.W. & Jurs, S. (1993). Using multiple regression to determine the number of factors to retain in factor analysis. Multiple Linear Regression Viewpoints, 20, 5-9.Bodie, Z., Kane, A. and Marcus, A.J. (2008). Investments. Seventh edition, McGraw-Hill, New York.