Generando y transfiriendo conocimiento en un entorno donde la única certeza es la incertidumbre

Libro de actas del IV Congreso Iberoamericano de Jóvenes Investigadores en Ciencias Económicas y Dirección de Empresas (AJICEDE)

Madrid, 16 y 17 de diciembre de 2021

Coordinadores/directores:
Dra. Gema Albort Morant
Dr. Antonio Luis Leal Rodríguez
Dr. Javier Morales Mediano
Dr. Ulpiano J. Vázquez Martínez

Generando y transfiriendo conocimiento en un entorno donde la única certeza es la incertidumbre

Edita, Red de Impresión

Coordinadores:
Gema Albert Morant
Antonio Luis Leal Rodríguez
Javier Morales Mediano
Ulpiano J. Vázquez Martínez

Depósito Legal: SE 226-2022
ISBN: 978-84-124399-41-5

Madrid, 16 y 17 de diciembre de 2021

IV CONGRESO IBEROAMERICANO AJICEDE

MADRID, 16-17 DICIEMBRE 2021

A COMPARISON STUDY BETWEEN COHERECE AND PERPLEXITY FOR DETERMINING THE NUMBER OF TOPICS IN PRACTITIONERS INTERVIEWS ANALYSIS

Lidia Pinto Gurdiel, Universidad Pontificia de Comillas, 201803276@alu.comillas.edu

Javier Morales Mediano, Universidad Pontificia de Comillas, jmorales@comillas.edu

Jenny Alexandra Cifuentes Quintero, Universidad Pontificia de Comillas, jacifuentes@icade.comillas.edu

Keywords: LDA, coherence, perplexity, topic modelling, text analysis

A COMPARISON STUDY BETWEEN COHERECE AND PERPLEXITY FOR DETERMINING THE NUMBER OF TOPICS IN PRACTITIONERS INTERVIEWS ANALYSIS

1.- INTRODUCTION

Over the last two years, the 90% of the currently existing data had been produced. Each minute, 15, 220,700 texts are sent, 3,607,080 Google searches are conducted and 456,000 tweets are sent, while the quintillion bytes of data generated on a daily basis keeps growing at an exponential rate (Schmidt, 2010). Large amounts of structured data are effortlessly being analysed using traditional software and major improvements have been made on the field. However, an increasing amount of this data is collected in an unstructured format, defying conventional methods for its study. Text mining methods are thus being developed in order to provide useful tools for the analysis of those inputs that have proven manual and traditional analysis to be unpractical (George et al., 2016).

Text mining is an automatic tool based on the natural language processing in order to obtain valuable insights from unstructured data, such as tweets, interviews, articles or even extense literary works. One of the most useful text mining methodology is topic modelling, which consists of the coding of the content of a text into meaningful subcategories called "topics" (Mohr and Bogdanov, 2013). The goal of a topic model analysis is to define a number of topics that accurately represent the text from which they have been extracted.

According to Lee et al. (2010), several models have been developed regarding topic modelling, being Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) the most widespread. The LSA is based on the projection of term-document matrix into a small factor space, achieving the reduction of the matrix dimension and its decomposition on smaller factors, the topics (Lee et al., 2010). The PLSA assumes that documents are generated by topics, and topics by words, and each child unit has a certain probability of belonging to one or other parent unit. Even though PLSA improves LSA by building a robust probabilistic theory, it does not fully reflect the generative process of a text at a document level. The LDA incorporates the whole process using a Dirichlet Distribution and has empirically proven to outperform the two previous methods when applied to texts composed by documents that mix several topics (Lee et al., 2010).

The Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003), is a generative probabilistic model for text corpus. Based on a three-level hierarchical Bayesian model, defines each document as a bag of words produced according to a mixture of topic probabilities. Each topic is in turn a distribution over all words found in the corpora, meaning that those words closely related with the document's main theme have a higher probability of being placed in its document's bag. Consequently, each document represents these topics in different proportions, as corpus tend to be heterogenous by combining several ideas that permeate the collection as a whole (Blei and Lafferty, 2009). The main objective of an LDA model is to select a number of topics that, combined, could accurately generate the original corpus. As LDA is an unsupervised method, topics are not predefined but learned by the model when associating words to topics according to their distribution. The number of topics, however, needs to be preselected, and several methods have been developed in order to choose the most adequate one on a general basis. Nonetheless, those methods may turn inefficient when applied to some specific documents. Over the last few years, multiple studies have been carried for extracting topics

from short texts, such as tweets or open-answer surveys. So, the corresponding use of techniques to select the number of topics in such texts is consolidated (Fang et al., 2016). However, to the best of our knowledge, no research has been carried out for the exploration of unstructured interviews to practitioners extracted from recognised and informative publications that deal with one specific business area.

In this paper, we investigate how the number of topics of a LDA model is defined based on the two most popular metrics: perplexity and coherence (Newman et al., 2010b) and compare the results of using each metric. To do so, we applied both algorithms to a collection of interviews about the Spanish private banking sector[1]. This paper is structured as follows; after a brief introduction in the first section, we will review the existing literature regarding the most common techniques to establish the number of topics. Afterwards, we will explain the methodology we have implemented in section 3 and present and analyse our results in section 4. Finally, in section 5, we will discuss the limitations of both metrics, as well as propose some other paths for further research.

2.- LITERATURE REVIEW

As mentioned before, electing the number of topics has an enormous effect on the displayed topics (Newman et al., 2010). There is a proven relationship between the number of topics and the probability of topics being nonsensical (Mimno et al., 2011). However, as the prime objective of topic modelling is to generate *k* number of topics so that their combination represents the whole original text, a too small size could keep us away from our primary goal. Consequently, one of the first steps when pursuing topic model is the election of the right number of topics. This choice is always done a priori, and two main metrics are available for carrying out this task: perplexity and coherence.

The most conventional evaluation of topic modelling consists of measuring the model predictive performance for unobserved documents. Mathematically, the perplexity score is inversely related to the model predictive likelihood, so a low perplexity score results in a better generalization performance (Bao and Datta, 2014). As the number of topics grow, the probability that the extracted topics from the training data cover all words from the original text decreases, as it is harder to generalize, and the perplexity grows. In other words, overfitting the model with multiple, unnecessary topics will only result in a bad predictive performance.

Even though the perplexity-based method achieves moderately good results, some researchers have questioned its stability, as results for the same dataset would vary depending on the chosen seed. Consequently, a new-perplexity approach was proposed: the Rate of Perplexity Change (RPC), which consists of extracting the average perplexity for every *k* number of topics (Zhao et al., 2015). Despite its mathematical basis, it is not clear whether the complexity of the model offsets the slights differences over the obtained results.

Both, perplexity and RPC are based on probability distributions, but do not reflect the topics semantic coherence, and some even suggest that, sometimes, it can be contrary to human judgements (Newman et al., 2010b). Moreover, models that only rely on these kinds of metrics are at risk of generating chained, unbalanced, intruded or random topics, as non-related words

[1] Private Banking refers to the specific services aimed at satisfying the financial needs of high-net-worth individuals (HNWI), usually delivered by the private banker (Morales Mediano and Ruiz-Alba, 2019).

may be included on the same topic. In order to avoid these errors, a metric based on the co-occurrence of words within the document is needed.

The most popular metric regarding semantic validation is coherence. Coherence rates topic quality based on human comprehension. A semantically coherent topic would be composed of a list of words which, collectively, are likely to represent a semantic theme (Mei et al. 2007). The scoring-based method relies on word co-occurrence statistics gathered either from the internal corpus or from an external one, such as Wikipedia. Recently, a new method based on combining internal and external corpus for topic coherence validation has been proposed. As vanilla coherence may generate basic topics, which, although semantically correct, might not fully represent the topics per se, the term frequency–inverse document frequency (tf-idf) coherence method privileges words that rarely occur in external corpus but do frequently occur in the internal corpus, generating discriminative topics (Nikolenko et al., 2017).

Even though several modifications over conventional metrics have been proposed, basic perplexity and coherence may be used for the comparison between semantic and numeric validation when electing the right number of topics.

3 - METHODOLOGY

The goal of our research is to discuss the existing metrics for assessing topic models. First, we carried out a structured review of the methodological approaches across the existing literature. We studied both coherence and perplexity metrics, as well as they proposed modified versions and then, we proceeded to apply them to our dataset. Firstly, we prepared our dataset by lemmatizing each word and removing stop words, such as articles and adverbs, so that they would not interfere in our study. coherence and perplexity metrics, as well as their proposed modified versions, we applied each method to the same dataset. The dataset used for this purpose consists of several interviews conducted by Fundspeople[2] to different Spanish executives about the Spanish private banking sector. Those interviews were pooled together into a unique document which, afterwards, was pre-processed; the document was split into sentences, punctuation and stopwords, words that add no information to the text such as determinants or pronouns, were removed. In addition, we eliminated as well some words which, despite their frequency, did not add any information to the text, such as “year” or “time” as well as company names.

Once the dataset was clean, we calculated the perplexity and coherence scores for each model containing a number of topics that ranged from 10 to 150, by tens, as:

$$perplexity\ score\ (D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\log_p(w_d)}{\sum_{d=1}^{M}N_d}\right\}$$

$$coherence\ score\ (D_{test}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D\left(v_m^{(t)},\ v_l^{(t)}\right)+1}{D\left(v_l^{(t)}\right)}$$

[2] FundsPeople is a leading community for collective investment and asset management professionals in Southern Europe (Spain, Andorra, Portugal and Italy). Aimed at fund management and distribution professionals, institutional investors and private banking professionals, summarises the main economic and financial news. More info at www.fundpeople.com.

Thereafter, two different models were parallelly generated, each one with the number of topics proposed by each metric. Then, we compared the obtained results by analysing topic quality and intertopic distances, which is based on the similarity between them. In this context, Figure 1 describes the structure of our analysis workflow.

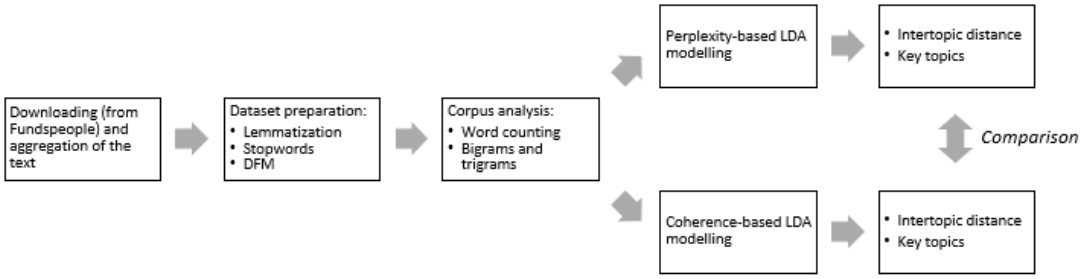

Figure 1: Workflow of study. Source: own elaboration.

4 - RESULTS AND FINDINGS

The dataset used for our study contains 35 interviews, conducted and collected in 2021, about the same subject: opinions on the evolution of the Spanish Private Banking after the COVID-19 pandemic. The final dataset, containing those interviews, constitutes a small corpus with multiple underlying subtopics which enabled us to apply different metrics in a small period of time. As mentioned before, the most repeated words in the text are likely to appear on one or another topic, and some of them will even belong to the same, depending on the chosen number of topics. Figure 2 shows the 15 most popular terms in descendant order, as well its frequency.

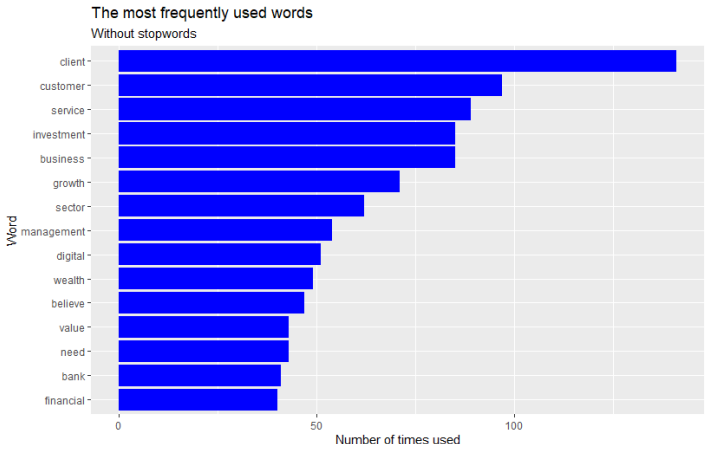

Figure 2: top 15 terms by frequency. Source: own elaboration.

Each individual term provides little insight about the text theme, even though they are clearly related to financial and economic topics (clients, investment, business…). Topics, optimally, consist of the combination of related words, words that are likely to co-occur throughout the text. Bigrams and trigrams, groups of 2 or 3 words that appear together several times, are also built, and the figures 3 and 4 show the most popular groups.

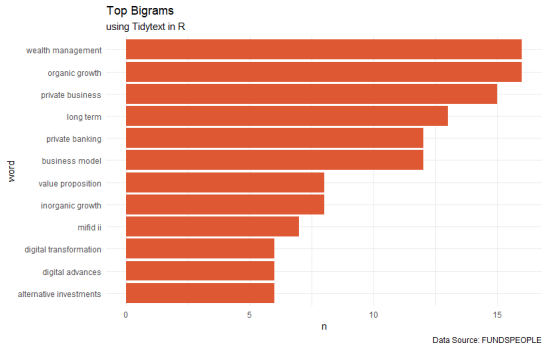

Figure 3: top bigrams by frequency. Source: own elaboration.

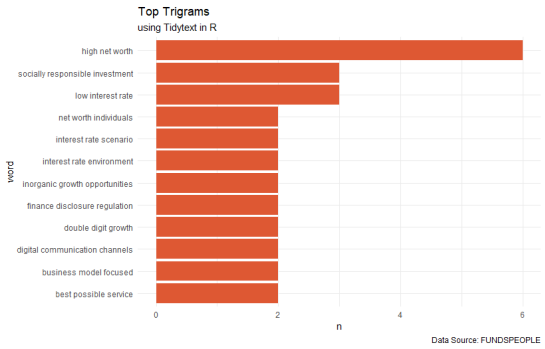

Figure 4: top trigrams by frequency. Source: own elaboration.

Despite their reduced lengths, these groups of words give us some insights about the essence of the text and, consequently, about the topics we will extract. Bigrams such as “wealth management” and “organic growth” and trigrams such as “high net worth” and “socially responsible investment” account for the most repeated groups of words used together throughout the text.

Afterwards, we created a document-feature matrix (dfm) and converted our dfm to a format capable of processing topic models. From now on, we followed two parallel different paths. The first path consists of applying perplexity for evaluating the model. For a list of possible number of topics, that ranges between 10 and 150, by tens, we generated a LDA and calculated its perplexity. The obtained perplexity score by each number of topics is returned and plotted, as in figure 5.

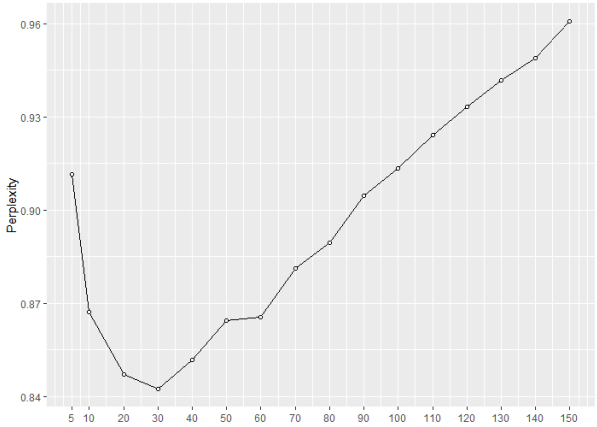

Figure 5: perplexity score for number of topics ranging from 5 to 150. Source: own elaboration.

In order to decide which number of topics is the most accurate by analysing this graph, the elbow method is applied. Therefore, we pick the k with the lowest perplexity score, hence we pick 30 topics. An LDA model with k=30 is processed. Meanwhile, another different approach is taken for choosing k. We carry out a similar process but the coherence score is calculated instead for each of the possible topics, taken from the same list of candidates. Results are plotted as in figure 6.

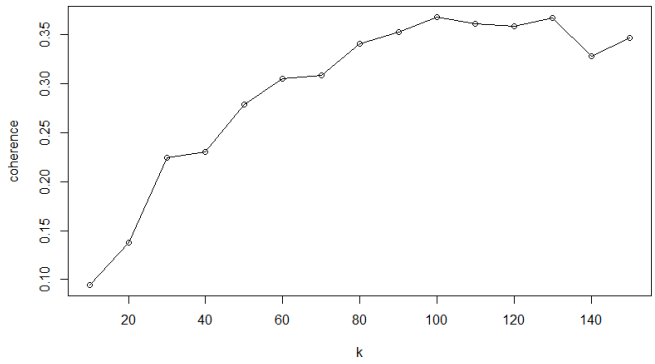

Figure 6: coherence score for number of topics ranging from 5 to 150.
Source: own elaboration.

In this case, the higher coherence is reached with around 100 topics. However, 100 topics would result in an overfitted LDA model. According to the elbow method, an optimal number of topics would be 60, as the marginal increase on the model's coherence decreases from then on. We could have picked 80 topics as well. However, considering the small size of our dataset, we considered that 80 topics would add too much complexity to our model. Moreover, the highest increase in the coherence score is observed within the first 60 topics. Therefore, we processed the second LDA model with k=60. Results for both numbers of topics are shown in the annexes.

As mentioned before, topics could be described as clusters of words. Those clusters are generated based on the number of topics elected and the corpus of words that constitute the text are therefore distributed in *k* number of topics which, together, represent the whole original text.

When *k* is too low, each topic is forced to represent a larger piece of the text, and the contained words share little semantic relation. As *k* increases, more clusters will be produced, and the more similar will be to each other. If the number of topics is too high, each additional topics will not provide additional information but overcomplicate the model.

In order to analyse whether the increase in topic quality offsets complexity, we picked the top 5 topics generated by each model. Selection was made on the basis of the percentage of total documents in which a topic appeared. For the perplexity-based model, topics 1, 3, 4, 2 and 9 where chosen, with frequencies of 10.9%, 7.1%, 7.0%, 6.5% and 5.2% respectively. For the coherence-based model, the elected topics where 1, 2, 3, 6 and 8, with frequencies of 7.0%, 6.8%, 4.7%, 3.8% and 3.6%. The top 10 words for each topic are shown in the tables 1 and 2.

Perplexity-based model	
Topic 1	Digital, wealth, situation, regulatory, level, service, solutions, tools, financial, needs.
Topic 3	Investments, wealth, value, financial, within, see, can.
Topic 4	Business, model, best, bank, social, good, needs, trust, banking, impact.
Topic 2	Decisions, significant, must, products, still, portfolios, particular, uncertainty, mind.
Topic 9	Organic, growth, crisis, last, still, services, challenge, online, understand, fundamental, economic, circumstances.

Table 1: top 5 topics resulted from the perplexity-based model. Source: own elaboration

Coherence-based model	
Topic 1	Wealth, financial, regulatory, products, importance, market, relationship, service, analysis, conservative.
Topic 2	Digital, clients, value, tools, trends, must, uncertainty, like, advisors.
Topic 3	Value, financial, able, decisions, client, important, particular, knowledge, must.
Topic 6	Investment, concentration, asset, future, technological, personal, traditional, regulated, profit.
Topic 8	Solutions, financial, relationship, sustainable, bankers, information, presence, security, standards, issues, generated.

Table 2: top 5 topics resulted from the coherence-based model. Source: own elaboration

Regarding the comparison of topics, we observed that perplexity-based methods produced simpler LDA models than coherence-based methods (30 topics versus 60 topics). However, despite its mathematical basis, perplexity proved to be a poor indicator of topic quality, as themes obtained were rather non-sense for human, since they usually do not share a logical connection. For instance, we consider topics 3, 4 and 5 from the perplexity-based results as remarkably difficult to interpret (see table 1).

Conversely, coherence-based notably optimized topic quality instead. Despite the higher number of topics, hence a more complex LDA model, humans will more likely be able to go through each topic and actually determine the theme which connects (almost) every word of a topic. We appreciated a much direct and easy interpretation in the five top topics offered by the coherence-based model (see table 2).

5 - LIMITATIONS AND FUTURE RESEARCH

The comparison that we have conducted regarding coherence and perplexity metrics was made on the basis of topic quality, as we aimed to determine which metric was able to generate topics which could be easily understood. Nevertheless, other variables such as the intertopic distance could provide further insights regarding similarity amongst topics in order to obtain more efficient models. On the other hand, the dataset used for our study is composed by informative interviews, which non-necessarily followed a common methodology. Even though these types of texts are the most common in an ordinary basis, analysis over a more structured text, obtained through a well-defined process, could produce different, and maybe better, results.

Taken all of this into account, it is clear that huge progress has been made in the field of text analysis. As the amount of unstructured data generated grows exponentially, new, effective tools are needed in order to extract information from that enormous source. Data analytics has always been based on quantitative tools, which makes these new inputs so hard to be analysed. Qualitative approaches are needed in order to fully understand the insights behind the data. Metrics such as coherence go one step beyond probabilities and start considering each topic as whole, which results in generating understandable topics for the final target: humans.

However, further progress is needed in the field. In order to be able to scale to even-larger datasets, an automatized method for choosing the right number of topics is needed, as well as a new metric. In smaller datasets, with less than a thousand rows, where not many different topics are found, coherence is an adequate metric for ensuring an acceptable range of comprehensible topics. However, in enormous datasets where hundreds of different themes are observed, coherence is likely to propose k number of topics which, although likely to lead to comprehensible topics, might not represent the essence of the text, or produce topics which are similar or even copies of others. Extracting k number of high-quality topics that actually represent the corpus should be pursued, and a new method combining perplexity and

coherence, as well as intertopic distance maximization could be a new field for further research, in order to generate coherent but non-overlapping topics.

Finally, further research is needed regarding topic quality. Even though coherence has achieved the introduction of qualitative criteria for topic assessment, topics are usually composed of general words which, despite their high-co-occurrence scores, are not discriminative. Thus, general topics which do not fully represent that concrete document are extracted.

BIBLIOGRAPHY

Bao, Y., and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371–1391.

Blei, D., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research, 3*, 993-1022.

Blei, D. M. and Lafferty, J.D. (2009). Topic models. In A. Srivastava, and M. Sahami (Ed.), *Text Mining: Classification, Clustering and Applications* (1st ed., pp. 71–93). Chapman and Hall/CRC.

Fang, A., Macdonald, C., Ounis, I. and Habel, P. (2016). Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data. In Ferro N. et al. (ed), *Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science* (1st ed., Vol. 9626, pp. 492–504). Springer, Cham.

George, G., Osigna, E.C., Lavie, D. and Scott, B.A. (2016). Big data and data science methods for management research. *Academy of Management Journal 59*(5), 1493-1507.

Lee, S., Baker, J., Song, J., and Wetherbe, J. C. (2010, May). An Empirical Comparison of Four Text Mining Methods. *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences, HICSS-43*, 1–10.

Mei, Q., Shen, X., and Zhai, C. (2007, August). Automatic Labeling of Multinomial Topic Models. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 490-499.

Mimno, D. M., Wallach, H. M., McCallum, A., Talley, E., and Leenders, M. (2011, July). Optimizing Semantic Coherence in Topic Model. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

Mohr, J. W. and Bogdanov, P. (2013). Introduction to Topic Models: What They Are and Why They Matter. *Poetics 41* (6), 545-569.

Morales Mediano, J. and Ruiz-Alba, J.L. (2019). New perspective on customer orientation of service employees: a conceptual framework. *The Service Industries Journal 39*(13-14), 966-982.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010a, June). Automatic Evaluation of Topic Coherence. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Association for Computational Linguistics*, 100-108.

Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T. (2010b, June). Evaluating Topic Models for Digital Libraries. *Proceedings of the Annual joint Conference on Digital libraries*, 215-224.

Nikolenko, S.I., Koltcov, S., and Koltsova, O. (2017). Topic Modelling for Qualitative Studies. *Journal of Information Science 43*(1), 88–102.

Schmidt, E. (2010, August). Every two days we create as much information as we did up to 2003. *Techonomy conference in Lake Tahoe, CA*.

Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modelling. *BMC Bioinformatics 16*(8), 1-10.