Scientometrics (2022) 127:5441–5458
https://doi.org/10.1007/s11192-022-04318-5

Validation of scientific topic models using graph analysis and corpus metadata

Manuel A. Vázquez[1] · **Jorge Pereira-Delgado**[1] · **Jesús Cid-Sueiro**[1] · **Jerónimo Arenas-García**[1]

Received: 8 April 2021 / Accepted: 17 February 2022 / Published online: 30 March 2022

Abstract

Probabilistic topic modeling algorithms like Latent Dirichlet Allocation (LDA) have become powerful tools for the analysis of large collections of documents (such as papers, projects, or funding applications) in science, technology an innovation (STI) policy design and monitoring. However, selecting an appropriate and stable topic model for a specific application (by adjusting the hyperparameters of the algorithm) is not a trivial problem. Common validation metrics like coherence or perplexity, which are focused on the quality of topics, are not a good fit in applications where the quality of the document similarity relations inferred from the topic model is especially relevant. Relying on graph analysis techniques, the aim of our work is to state a new methodology for the selection of hyperparameters which is specifically oriented to optimize the similarity metrics emanating from the topic model. In order to do this, we propose two graph metrics: the first measures the variability of the similarity graphs that result from different runs of the algorithm for a fixed value of the hyperparameters, while the second metric measures the alignment between the graph derived from the LDA model and another obtained using metadata available for the corresponding corpus. Through experiments on various corpora related to STI, it is shown that the proposed metrics provide relevant indicators to select the number of topics and build persistent topic models that are consistent with the metadata. Their use, which can be extended to other topic models beyond LDA, could facilitate the systematic adoption of this kind of techniques in STI policy analysis and design.

Keywords Topic modeling · Latent Dirichlet Allocation · Graph analysis · Semantic similarity · Model validation

✉ Jerónimo Arenas-García
jarenas@ing.uc3m.es

Manuel A. Vázquez
mavazque@ing.uc3m.es

Jorge Pereira-Delgado
jpereira@ing.uc3m.es

Jesús Cid-Sueiro
jcid@ing.uc3m.es

[1] Universidad Carlos III de Madrid, Madrid, Spain

Introduction

The growing interest in methods based on natural language processing (NLP) and machine learning has driven intense research work for its application in the field of science, technology and innovation (STI) analysis. In this work we focus on probabilistic topic modeling, and more specifically on latent dirichlet allocation (LDA) (Blei et al. 2003; Srivastava and Sutton 2017; Xiao and Stibor 2010), a machine algorithm that allows the identification of the predominant topics in a corpus of text documents.

LDA takes as inputs the natural language texts from the corpus, and provides the following two outputs:

1. A list of predominant topics, where each topic is defined by a weighted list of characteristic terms.
2. A mapping of documents from the word space into a topic space, in which each document is characterized by a vector of length equal to the number of topics, each component yielding the proportion of the document that is assigned to its corresponding topic.

With respect to the widespread use of statistical approaches supported by taxonomies, LDA has some properties that are useful for STI analysis. Among them are:

– Thematic analysis can be carried out with different levels of resolution, or hierarchically. This helps in the survey of specific subject areas with the desired level of detail.
– Its flexibility allows the identification of emerging topics, or the detection of hybridization of topics, something hard to do using taxonomies that are often exclusive and whose updating imposes certain time delays.
– Since LDA also provides a vector representation of documents, it is possible to carry out a semantic comparison between documents of different data collections, even in the absence of a common taxonomic representation. Also, the time shift of certain topics in one corpus with respect to another can be studied (lead-lag analysis).

These properties have motivated recent works that exploit the modeling of topics on scientific corpora (Suominen and Toivanen 2016; Boyack et al. 2011; Colavizza et al. 2021) (typically research articles, patents, and funding applications in the form of project summaries), and even platforms aimed at policy makers that provide valuable information for science analysis, tracking and policy decision making. As an example, Corpus Viewer (Pérez-Fernández et al., 2019)[1] is a platform developed by Spanish Government institutions used for the evaluation of research proposals and the analysis of scientific activity in Spain. In the European context, Data4Impact[2] is a project that seeks to measure the economic and social impact of public funding in the field of health, while Arloesiadur[3] is a Welsh initiative to model topics and detect trends in industry and research. These and other examples of systems for the *digitalization of science and innovation policy* can be found in (OECD, 2018).

[1] https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/plataformas/Paginas/corpus-viewer.aspx
[2] https://cordis.europa.eu/project/id/770531/es
[3] https://arloesiadur.org.

Despite these successful applications, some well-known practical drawbacks hinder a more generalized adoption of these tools. First, the use of LDA requires the adjustment of hyperparameters that have a significant impact on the model performance: mainly, the number of topics, whose selection may be oriented to satisfy the preferences of the end user, and the concentration hyperparameters, which have a major effect on the nature of the topics obtained (whether they are more general or more specific, i.e., characterized by a greater or lesser number of terms), or on the assignment of documents to topics (whether a document encompasses a small or large number of topics). Furthermore, the stochastic components in the available implementations of LDA provide non-deterministic results: different runs of the algorithm yield different topics. This variability, which can be highly dependent on the hyperparameter values, reduces the reproducibility of the results (Lancichinetti et al. 2015) and their interpretability, and harms the confidence of the end user in the technology (Grant et al. 2017).

The most commonly used approaches so far to adjust the LDA hyperparameters are based on the use of log-perplexity (Blei and Lafferty 2006) and different measures of topic coherence (Newman et al. 2010). However, high log-perplexity does not guarantee low variability, and some studies show that these measures are not always well correlated with human interpretability (Chang et al. 2009).

These and other metrics for hyperparameter selection are focused on the quality of topics. However, in STI applications, the quality of the document similarity measures derived from the topic model are of the utmost importance, because these are the source for ulterior analyses (document clustering, impact indicators, etc). Our work is motivated by the need to define validation procedures that are centered on the metrics defined in the LDA vector space, and not on the topic descriptions.

Therefore, our goal in this paper is to introduce and assess a new methodology for the validation of topic models. It relies on two metrics that can be used for tuning and/or validating the hyperparameters that one must set in training a topic model. These metrics are specifically designed to minimize variability and favor correlation of the models with other metadata which is usually available for STI datasets, typically citations or category labels related to scientific discipline, applications, etc. Thus, unlike existing criteria, our main interest lies in the usability of the topic model from the point of view of the users to whom it will be delivered. Ultimately, by providing more persistent models which are also coherent with existing metadata, we pursue to increase users' confidence and favor the adoption of topic models as a tool for STI analysis and other policy tasks.

The two proposed metrics rely on the idea that the quality of a topic model is given not only by that of the topics returned by the algorithm, but also by the quality of the vector representation of the documents, which can be evaluated through the analysis of the document similarities derived from the LDA vector space.

We study the performance of the proposed metrics and their correlation with the usual criteria of coherence. In particular:

– For the stability measure, the variability of the distances between pairs of documents is analyzed. This allows to obviate differences between models that are simply due to topic permutations or varying composition in terms of low relevance.

– For the alignment with other available metadata, the similarity between the *semantic graph* constructed from the topic models and a *reference graph* obtained from the metadata is measured.[4]

A major issue of the proposed metrics is that they imply inter-document similarity calculation, which can be a very costly task for large datasets. However, despite the fact that the size of the graphs grows quadratically with the number of documents, it is possible to work with large graphs by putting bounds on the semantic distances and distributing the calculation over Graphic Processing Units (GPUs) that allow for efficient parallelization. Furthermore, the graphs analyzed are typically sparse (the density of links is low), and the study can even be restricted to the documents that exhibit the largest similarity.

The suitability of our approach is illustrated in the experiments section by first analyzing the proposed metrics on a synthetic dataset for which the *ground-truth* hyperparameters are known. After that, we devote our efforts to more challenging scenarios using three real datasets from the STI domain: scientific papers, patent applications, and funding proposals. Our experiments show that the proposed metrics are useful on their own, and can be combined with coherence to select the most critical hyperparameters of the LDA algorithm.

The rest of the paper is organized as follows. In the next section, previous work regarding the validation of topic models is reviewed. Sections 3 presents our approach for LDA model validation based on the two proposed metrics. Sections 4 and 5 are dedicated, respectively, to the description of the experimental setup and discussion of the results. Finally, Section 6 concludes the paper and identifies some lines for future research.

Related work

Since the publication of its seminal paper (Blei et al. 2003) in 2003, LDA has been one of the most prevalent methods aimed at the analysis of large corpora of documents. This statement is backed by the number of works based on LDA that have been published in the last few years in fields such as information retrieval (Han 2020; Miyata et al. 2020), library and information science (Adebiyi et al. 2019; Xue 2019), or scientometrics (Burghardt and Luhmann 2021; Ranaei et al. 2020). However, very frequently, they ignore or gloss over the issues of validation and/or hyperparameter tuning.

There have been several attempts to try and tune the main parameters of LDA in a systematic way that leaves out the human factor. For instance, differential evolution is used in Agrawal et al. (2018) to find an appropriate number of topics, as well as prior distributions, so that LDA yields consistent results. In order to quantify this, an *ad-hoc* stability metric is proposed. In the same vein, simulated annealing serves in Pathik and Shukla (2020) to (separately) tune the priors' parameters and the number of topics. A completely different heuristical approach is proposed in Zhao et al. (2015), where the authors exploit the rate of change of the perplexity as a function of the number of topics to try and select an optimal value for the latter.

As already mentioned, another important issue when using topic models is that different runs of the algorithm (after fixing the hyperparameters) can produce significantly different

[4] Note that in this paper the term *semantic graph* is used to refer to a graph in which the nodes are documents and the edges represent document similarity calculated from the topic representation of the document.

results. In Mantyla et al. (2018) LDA is run a number of times and the topics from all the runs are afterwards clustered together to compute different metrics of stability. A similar idea is explored in Vega-Carrasco et al. (2020) through the use of a collection of *posterior* samples of the topic model provided by an *ad-hoc* implementation of LDA based on Markov Chain Monte-Carlo. The instability of topic models is also the main concern in Chuang et al. (2015), where a visualization tool is introduced to decide (through human intervention) which topics remain stable across different runs of the same (exact) topic model. Ultimately, the authors conclude that "a single topic model may not capture all perspectives on a dataset". A different strategy, also relying on subjective judgement, is put forward in Syed and Spruit (2017), where a human-built topic ranking is compared against coherence scores. While all these constitute sensible approaches in validating topic models, they are still unsuitable in real-world scenarios involving huge corpora that can hardly be analyzed manually.

In addition to the aforementioned problems, and closely related to the choice of hyperparameters and the non-deterministic nature of the LDA algorithm, we also have to decide whether the end product of the algorithm is good enough for our purposes. Validating a topic model ultimately entails making sure its output and/or the conclusions derived from it are consistent with some facts that have been established beforehand. This is especially true when the topic model is to be exploited together with other available metadata. Indeed, even though most strategies for validating topic models involve human intervention of one kind or another, external data sources can also be used if available. The authors of Hagen (2018) identify three possible ways to evaluate a topic model, two of which rely on human intervention. The third one is based on comparing the results given by the topic model with those obtained from another corpus of documents using simple (not based on topic models) techniques. A validation methodology for tuning hyperparameters in order to obtain accurate topics with low variability (i.e. high reliability) is proposed in Maier et al. (2018). The variability (measured as reproducibility) and the validity of topics is also the main concern in Hecking and Leydesdorff (2019) However, in both cases the validation is focused on the quality of topics, while the quality of the document similarities is not included in the methodology.

Our approach in this paper, which relies on the comparison of the document similarity graph (the *semantic* graph) with a reference graph is in some sense dual of that in Waltman et al. (2020), which employs text-based similarity measures to validate document clustering algorithms exploiting different types of citation-based metrics. The use of citations as a reference to evaluate texts similarity by means of topic models has also been proposed in Chen et al. (2020).

Proposed metrics for LDA model validation

Since topic modeling algorithms are essentially unsupervised, there is not a ground-truth allowing to evaluate the quality of the topics obtained after a single execution of the algorithm. Even if there are some categories or keywords associated with documents in the corpus that could be potentially useful to evaluate the topics, matching the categories to the topics is not straightforward, and would require manual intervention.

The unsupervised nature of LDA is an intrinsic property of the algorithm: the model is expected to discover the topic structure, without being anchored to a standard category set.

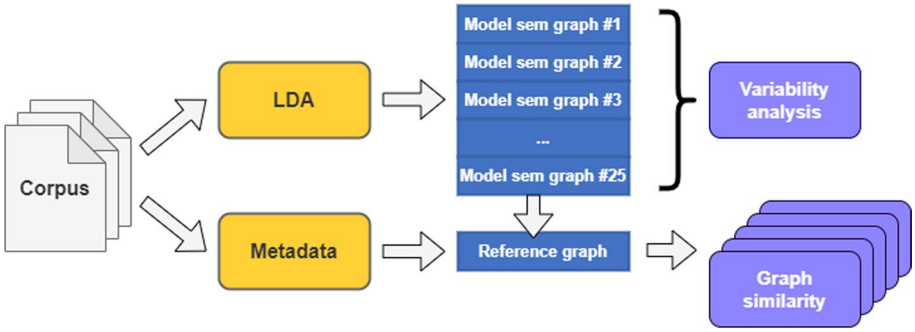

Fig. 1 Validation schema. For each combination of hyperparameters *N* LDA models are trained, and their corresponding semantic graphs calculated afterwards, in order to measure the variability. Additionally, each semantic graph is compared to a reference graph producing *N* estimated graph similarity values that are averaged to produce the final estimate

The use of category labels or keywords may bias the results of the model, reducing some of its potential utilities.

Our work is based on the idea that the quality of a topic model can be evaluated by its utility to capture the thematic similarities between documents. Thus, we are not interested in the specific topic-vector representation of documents. From our point of view, the intrinsic variability of a topic modeling algorithm would be irrelevant if all models generated by different executions of the algorithm apprehended the same document similarity relationships. This fact suggests a simple procedure to evaluate a topic model: compute a graph with nodes representing documents and edge weights given by inter-document similarities, and evaluate the quality of the topic model based on that of the similarity relations in the graph. In this work, such graph will be referred to as the *semantic graph*. Since we are mostly interested in evaluating the capability of the models to identify similar documents in the collection, we will ignore small similarity values and make semantic graphs sparse by removing edges with weights below a predefined threshold.

We only consider the validation of the number of topics and the document *a priori* distribution of LDA, though other hyperparameters could also be selected following a similar approach. The overall validation process considered here has been summarized in Fig. 1 and consists of the following steps:

1. For each hyperparameter combination *N* LDA models are trained which, in turn, are used to compute *N* semantic graphs using the corresponding document representations given by each one.
2. Variability analysis is based on the *N* semantic graphs computed for the same combination of hyperparameters using the proposed metric for graph variability.
3. A reference graph is computed from available metadata (using, e.g., citations or category labels), and its similarity with the semantic graphs used to compute the *graph similarity* metric.

The rest of this section is devoted to explain how the semantic graphs are computed in a memory and computationally efficient way (step 1), and how to use them afterwards to implement the validation metrics (steps 2 and 3). Note that the variability analysis is completely general and can be carried out on any available corpus, whereas the metric in step 3

requires a reference graph. However, for most commonly used STI datasets such reference graph can be computed from available metadata, as explained later.

Semantic graph computation

Topic-vector representations of documents given by the LDA algorithm correspond to a probability distribution over the topics, i.e., all the components in the vector representation of the i-th document, $\mathbf{x}_i$, lie in the interval $[0, 1]$, and altogether add up to 1. Therefore, the similarity between any two documents can be computed using common distance or divergence measures between probability distributions. The semantic graph is then computed from the topic model by taking the documents as nodes and the similarity values as edge weights.

For the sake of computational efficiency, similarities based on the Hellinger distance are used, i.e. if $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the topic representations for documents i and j, respectively, according to a particular LDA model, their semantic similarity will be computed as

$$w_{ij} = 1 - H^2(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\mathbf{x}_i}^{\top} \sqrt{\mathbf{x}_j}, \tag{1}$$

where $H(\mathbf{x}_i, \mathbf{x}_j)$ is the Hellinger distance between the two vectors, and $\sqrt{\mathbf{x}}$ is obtained by calculating the square root of vector $\mathbf{x}$ component-wise. This similarity measure is also known as the Bhattacharyya coefficient between probability vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

Since w_{ij} needs to be computed for every pair of documents, we can do so in an efficient way as

$$\mathbf{W} = \sqrt{\mathbf{X}} \sqrt{\mathbf{X}}^{\top}, \tag{2}$$

where $\mathbf{W}$ is the adjacency matrix of the semantic graph, $\mathbf{X}$ is the document-topic matrix returned by the LDA algorithm (with document representations $\mathbf{x}_i$ arranged row-wise), and the square root being computed element-wise. Notice that even though the memory and computation requirements for obtaining $\mathbf{W}$ grow quadratically with the number of documents, the whole process is amenable to be parallelized.

By exploting the sparse nature of $\mathbf{X}$[5], it is possible to further speed up the computation of $\mathbf{W}$ which, in turn, can greatly reduce the time required for finding the similarities between documents since one only needs to consider the overlapping non-zero components (Badenes-Olmedo et al. 2020). The sparsity in X induces also some sparsity in $\mathbf{W}$, but it might not be enough and, in general, all similarity values below a given threshold, w_{th}, are also set to zero. As a result, the weight matrix $\mathbf{W}$ becomes even more sparse. There are several reasons to enforce sparsity:

- Computation and memory constraints: the validation process may require the computation and storage of a considerable number of large graphs. Thresholding contributes to alleviate this problem.
- In practice, the computation and memory requirements can be reduced by only performing validation on a subset of documents from the corpus. In that case, thresholding

[5] In LDA, the number of *active* topics for each document (non-zero elements in the corresponding vector, $\mathbf{x}_i$) depends on the *a priori* distribution used to generate documents representation, and is typically rather small.

may be not strictly necessary. However, in applications where, after the validation, a large semantic graph will be computed for large corpora (on the order of millions of nodes) sparsity becomes unavoidable. Thus, carrying out the validation process using graphs with similar sparsity to that of the final graphs is a sensible approach.
– Since the reference graph will be computed from a different data source (metadata), thresholding serves to align all graphs to the same degree of sparsity. This is useful to avoid any bias induced by the original sparsity in the reference graph.
– In general, we are interested in identifying documents that are similar. Thresholding helps to focus the validation metric in the capability of models to identify similar documents, and ignore differences in small similarity values.

Given the above considerations, in our experiments the threshold w_{th} has been selected in such a way that, for each corpus, the sparsity degree of all semantic graphs and the reference graph is the same.

Metric 1: semantic graph variability analysis

The first metric used for model validation aims at measuring the inherent variability of the N semantic graphs inferred for each combination of hyperparameter values. For graphs obtained from topic models, this variability may originate from the stochastic nature of the non-convex optimization algorithms used to fit the models.

The proposed metric is simply the average of the standard deviations of the observed similarities among every pair of documents, i.e.,

$$V = \frac{1}{D^2} \sum_{i,j} \sqrt{\frac{1}{N} \sum_{n} (w_{ij}^{(n)} - \bar{w}_{ij})^2}, \tag{3}$$

where $w_{ij}^{(n)}$ is the semantic similarity between the i-th and j-th documents for the nth model, $\bar{w}_{ij}$ is the mean of these values over all semantic graphs, N is the number of trained models, and D is the number of documents (i.e. the number of nodes in the graph).

Metric 2: alignment with a reference graph

If a reference graph providing a ground-truth is available, the quality of a semantic graph can be estimated by comparison with such ground-truth using, e.g., the cosine similarity. Hence, the *reference graph similarity* (RGS) is defined as

$$\text{RGS} = \text{sim}(\mathbf{W}, \mathbf{R}) = \frac{\text{trace}(\mathbf{W}^\top \mathbf{R})}{\sqrt{\text{trace}(\mathbf{W}^\top \mathbf{W})\text{trace}(\mathbf{R}^\top \mathbf{R})}}, \tag{4}$$

where $\mathbf{W}$ and $\mathbf{R}$ are the adjacency matrices of the semantic and the reference graph, respectively.

In general, a ground-truth reference graph is not available (indeed, for our experimental work this information is only available for the synthetic data), but it can be replaced with an approximation based on document metadata. In the next section, the experimental setup to evaluate the suitability of the proposed metrics is described, explaining which reference graphs were obtained for each dataset used.

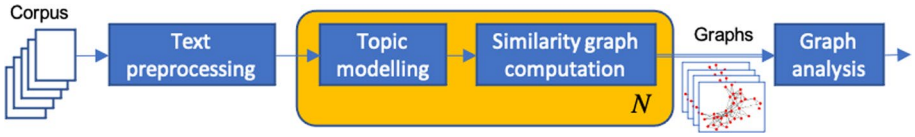

Fig. 2 Block diagram illustrating the processing steps of our experimental setup. Text preprocessing is initially done for each corpus to prepare it for topic modeling. Then, topic models and their corresponding similarity graphs are computed N times for each hyperparameter combination considered. Finally, model validation is carried out through graph analysis using the proposed metrics

Experimental setup

In order to analyze the performance of the proposed metrics for topic model validation, we have carried out experiments using synthetic and real corpora. For the synthetic data, LDA hyperparameters are known, as well as a ground-truth reference graph. To analyze the performance of the metrics in more realistic scenarios, we have included experiments using three real-world corpora from the STI domain: scientific papers, patent applications, and project summaries.

Fig. 2 shows the block diagram of our experimental setup. The whole process consists of 4 main steps: (1) a standard text processing block to get the *bag of words* representation of each document, (2) topic modeling, to project the documents onto a semantic space, (3) semantic graph computation, and (4) model validation using the two metrics proposed in the previous subsections.

Notice that, as explained in the previous section, topic models and their associated semantic graphs are computed N times for each combination of hyperparameters to be tested, so that model (and, therefore, graph) variability is also accounted for. During the final steps, graph metrics are computed in order to evaluate the models and select the best set of hyperparameters.

The following subsections describe in more detail the different components of our experimental setup, starting with an account of the datasets employed.

Dataset description

In order to test the proposed validation metrics under controlled conditions, a synthetic dataset will be used. Documents therein have been generated using a vocabulary of 50,000 terms with no semantic meaning (e.g., *term317*, *term56314*) from a given LDA generative model. The corpus consists of 100,000 documents, each of them with a random length drawn uniformly in the range [150, 250]. The true number of topics in the LDA model is 50, whereas the prior Dirichlet parameters are $\alpha_s = 0.1$, for the documents, and $\beta = 0.01$, for the topics.

In order to study the performance of the metrics in more realistic scenarios, three different STI text collections are considered, each of them with different types of metadata. The first corpus consists of a subset of 200,000 paper abstracts, belonging to the Semantic Scholar (S2) (Ammar et al. 2018) database, and more specifically, to a subset of papers which also belong to the PubMed (Wheeler et al. 2005) database. The second corpus has been downloaded from the National Institutes of Health (NIH) ExPORTER (Wheeler et al. 2005) catalogue and consists of 100,000 project summaries. The last

Table 1 Main features of the datasets used in the experiments.

Dataset	Type of documents	# docs	Vocab. size	Average doc. size
Synthetic	Synthetic data	100,000	9,826	199.36
Semantic Scholar (S2)	Scientific papers	200,000	150,498	120.75
PATSTAT	Patent applications	199,993	19,942	84.87
NIH RePORTER	Project summaries	99,925	25,575	171.23

corpus is a subset of 200,000 patent applications belonging to the PATSTAT (EPO, 2020) database.

Text preprocessing

Paper abstracts, patent applications and project summaries are passed through a series of NLP pipelines for English language. Text preprocessing components include tokenization, lemmatization, N-gram identification and stopword removal. For this, IXA pipes library (Agerri et al. 2014) has been exploited through a docker service that allows parallel execution (Badenes-Olmedo et al. 2017). In addition to this, we have created an additional corpus-specific stopword list, including tokens that appear either too often (in more than 60% of the documents) or too rarely (in less than 10 documents) in the dataset. As a result, the S2 paper abstracts corpus has a vocabulary size of 150,498 tokens, the NIH project summaries corpus has 25,575 tokens and the PATSTAT patent applications corpus has a vocabuarly size of 19,942 tokens.

Likewise, after carrying out the previous steps, a number of documents have been removed from the datasets due to insufficient length or the abundance of corrupted characters for which a unicode equivalent could not be easily obtained. As a result, the final S2, PATSTAT and NIH corpora used to train the models consisted of 200,000, 199,993, and 99,925 documents, respectively. Table 1 collects summary information about each dataset.

Topic modelling and hyperparameter exploration

In our experiments, we have used the implementation of the LDA algorithm provided by the Mallet library McCallum (2002), which is based on Gibbs sampling Yao et al. (2009). The following LDA parameters have been explored:

- Number of topics: 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150.
- Concentration hyperparameter of the (symmetric) Dirichlet distribution for document weights: $\alpha \in \{0.1, 0.5, 1, 5, 10, 20, 50\}$. The use of an asymmetric prior was also explored by enabling parameter re-estimation every 10 iterations.

Regarding the concentration hyperparameter of the Dirichlet distribution describing the topics, the default value provided by the library, $\beta = 0.01$, was used.

For validation purposes, $N = 25$ models per combination of hyperparameters are generated, for a total of 2,100 models per corpus. However, the observed variances on the metrics suggest that a smaller number of models (around $N = 10$) might be enough in most cases.

As briefly outlined in the introduction, each LDA model provides two relevant outputs:

- The description of the topics found, consisting essentially of, for each topic, a different probability distribution over all terms in the vocabulary. Topic description can be used to evaluate topic coherence Syed and Spruit (2017), that can then be averaged over all topics in the model, and over the N models for each hyperparameter combination. This will be used for comparison with our approach.
- Topic-vector representations for each document, i.e., the vectors $\mathbf{x}_i$ used to compute the semantic graph associated with each model.

Reference graphs

In order to calculate the second metric that has been proposed, a reference graph is required for each corpus. For the synthetic data, the ground-truth vector representation of each document is available and, therefore, a ground-truth reference graph can be straightforwardly obtained. Regarding the real-world datasets, no ground-truth is available, so we need to rely on the available metadata to generate the reference graphs.

For the S2 corpus, the bibliographic coupling (i.e. the relative amount of common bibliographic references) between papers has been used. Citation networks have proven to be useful sources for topic analysis [see Small et al. (2014), for instance], and it has been shown that bibliographic coupling provides better relatedness measures for document clustering than other citation-based measures like the direct citations or the co-citations (Waltman et al. 2020). To be more specific, $\mathbf{R}$ is in this case the adjacency matrix of a graph of bibliographic couplings obtained using citations in the S2 database. Its entries are calculated through the Jaccard index based on the sets of citations from each paper,

$$r_{ij} = \frac{|\mathcal{C}_i \cap \mathcal{C}_j|}{|\mathcal{C}_i \cup \mathcal{C}_j|}, \tag{5}$$

where $\mathcal{C}_i$ and $\mathcal{C}_j$ are the sets of citations from papers i and j, respectively.

In the case of the NIH corpus, consisting of project summaries, we have used the project keywords available for each project and compared them in a similar way, i.e., the similarity between two projects included in the reference graph is the Jaccard index based on the sets of keywords for each project.

Finally, for the PATSTAT corpus the metadata used is the *techn_field* value from table *tls230_appln_techn_field* of the PATSTAT database. This taxonomy divides the patents in 35 categories and assigns, for each patent, a vector of weights adding up to 1. The similarity between patent applications is measured based on that between these vectors. More specifically, matrix $\mathbf{R}$ was computed by means of the pairwise dot product of the *techn_field* vectors available in PATSTAT.

Recall that, ultimately, we aim at working with sparse graphs, and for that reason edges whose similarity is below a given threshold, w_{th}, are dropped. Here w_{th} is set so that the average number of edges per node is 100.

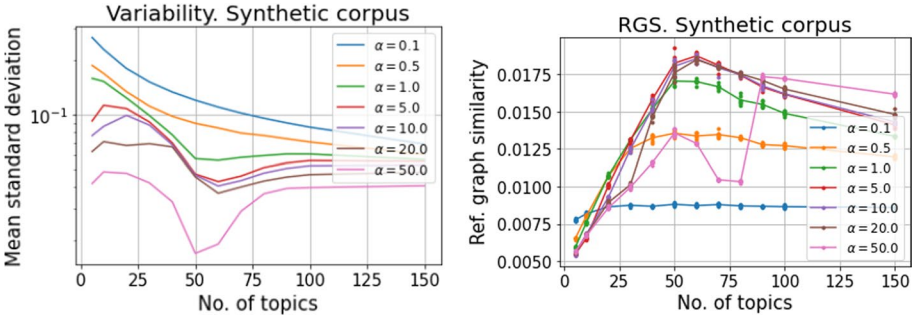

Fig. 3 Variability and reference graph similarity (RGS) calculated over the synthetic dataset. For the RGS metric (right) the plot shows the average over the trained models, as well as its individual value for each of the N models

Results

In this section we show the results of applying our method to each corpus. We start analyzing the proposed metrics on the synthetic dataset and move on, afterwards, to describe their performance on the real datasets.

Synthetic dataset

As explained in the methodology, for each combination of the two hyperparameters being explored, i.e., the number of topics and the concentration parameter α, $N = 25$ LDA models have been trained. Fig. 3 illustrates the computed variability metric (left subplot) and average reference graph similarity (RGS, right subplot). For the RGS subplot, in addition to the average RGS over the 25 trained models, the individual value for each one of them is also represented.

Focusing on the variability metric V given by Equation (3), one can see that, for most values of α, it presents a minimum around 60 topics (50 topics for $\alpha = 50$), very close to the true number of topics used to generate the dataset. This suggests that, in addition to a desired model property, low variability is a useful criterion for selecting the number of topics for the model. It is worth noting that the variability decreases as α increases (see, e.g. $\alpha = 50$), although the RGS can degrade significantly. This effect may be due to the fact that large values of α favor document representations with a lot of active topics, and increasing the overlap among active components decreases the variability. However, this *flattening* of the document representations may have a negative effect from the perspective of RGS.

In the RGS subplot it can be seen that intermediate values of α yield the largest similarity, including the theoretical value $\alpha = 5$[6]. It is observed that the maximum is also obtained around 60 topics.

All things considered, we can conclude that the joint analysis of both metrics would allow the selection of a combination of hyperparameters very close to that used for the

[6] As explained when describing the generation of the synthetic dataset, the optimum value of α_s is 0.1, but the Mallet toolbox takes the given α and divides it by the number of topics. Hence, a value of $\alpha = 5$ in Mallet is equivalent to $\alpha_s = 0.1$ for 50 topics.

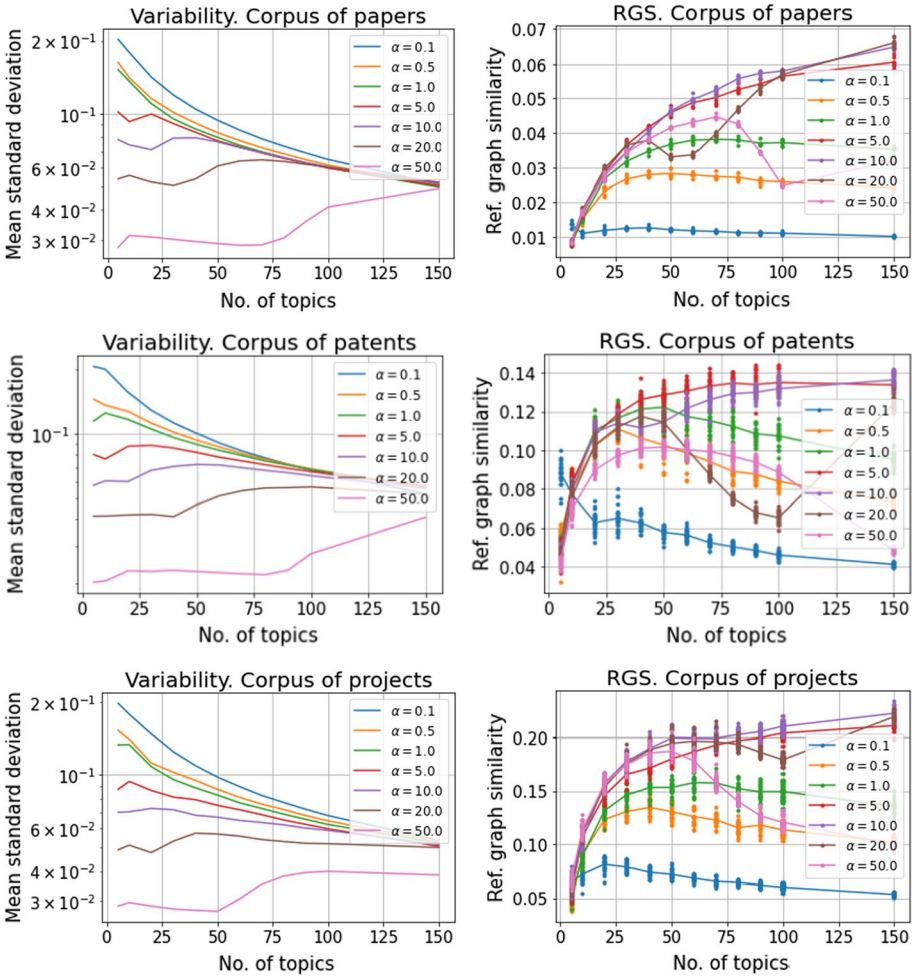

Fig. 4 Variability of the topic models and similarity with respect to the reference graph (RGS) of the three real datasets depending on the value of α and the number of topics

generation of the dataset. Both criteria approximately agree on the number of topics and show large RGS and stable variability in the range $5 < \alpha < 20$. Our suggestion in this respect would be to select the smallest α with these properties to favor sparser document representations.

Real-world datasets

Next, we move on to the analysis of results in the three real datasets. In this case, as documents were not actually generated by the LDA model, and hence a ground-truth is not available, we need to discuss our results from a qualitative point of view. Variability and RGS are displayed for the three corpora in Fig. 4

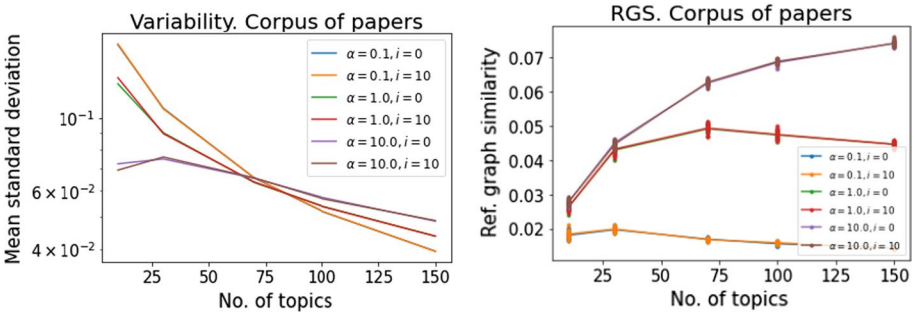

Fig. 5 Semantic Graph variability (left) and Reference graph similarity (RGS) with and without re-estimation of word-document distribution parameters for the S2 corpus. The legend indicates the initial value used for the α parameter ($\alpha \in \{0.1, 1, 5\}$) and whether parameters were fixed ($i = 0$) or reestimated every 10 iterations ($i = 10$). For the latter case, the LDA implementation will learn an asymmetric prior with parameters reestimated every 10 iterations

The right panels of Fig. 4 show the plots of the RGS, one per corpus, as a function of the number of topics and the values of the hyperparameter α. Every plot illustrates that the value of α has a large impact on the quality of the graph. The optimal value of α depends on the number of topics and the corpus but, in general, intermediate values around 5 and 10 yield the highest RGS.

For projects and patents, and after dismissing extreme values of α, it can be observed that the RGS grows up until around 50 topics, and stabilizes above that number, especially for the best values of α. This suggests that, at least for these corpora, the number of topics should be at least 50, (though the specific value can be selected attending to other criteria, like the variability or the coherence). Finally, for scientific papers the RGS keeps growing up until at least 150 topics, though the increment is relatively small from 70 topics onwards.

Regarding the variability of the topic models (left panels in Fig. 4), we can see that, in general, larger values of α have smaller variability, although for the range of values of interest according to the analysis of the RGS, the influence of α on the variability is smaller. This suggests that the selection of α should be guided mainly by the RGS (same conclusion as before for the synthetic data).

Interestingly, the local minima of V around the true number of topics that was observed for the synthetic data can also be observed for the real datasets, though it is less stressed, and its position less stable with respect to α. Although the location of the minimum seems to underestimate the best values of the RGS according to the number of topics, we can see that, even though $\alpha = 50$ did not provide good RGS values, it shows the deepest valley of the variability, which is in good agreement with the best values of the number of topics obtained through the RGS analysis. A further study of the variability seems convenient to obtain a more robust estimator of the number of topics, for which we will explore different normalization strategies to account for the differences observed in the document similarity distributions when varying the number of topics (distributions skew towards 0 when increasing the number of topics).

The role of asymmetric priors for the Dirichlet distribution responsible for document generation has also been explored, since this is a setting that some authors have reported to provide advantages. In order to analyze this, we enabled Mallet reestimation of the *a priori* parameters for such prior distribution, which estimates a different α for each topic. Fig. 5

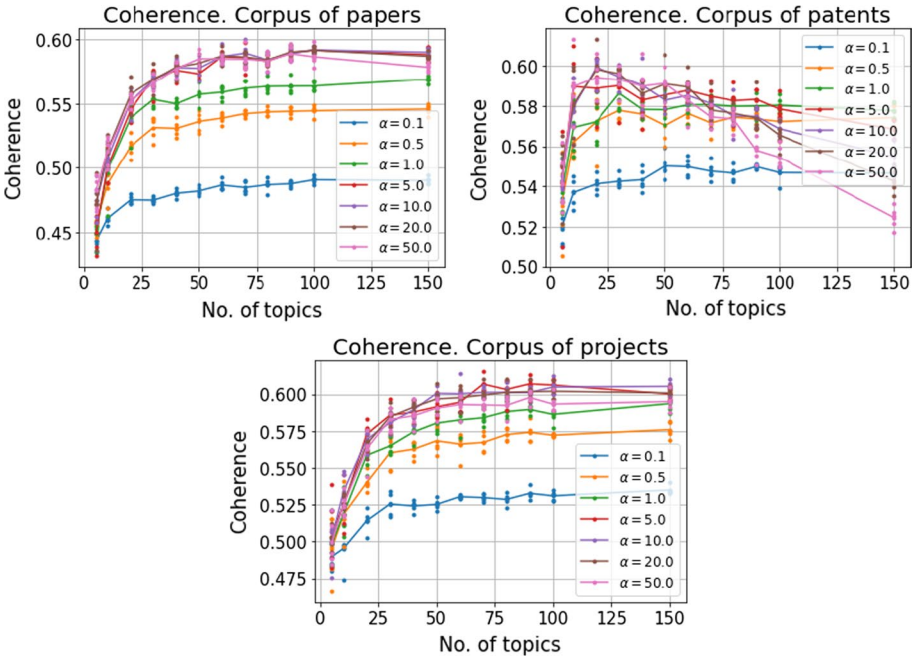

Fig. 6 Coherence of the topic models with respect to the value of alpha and the number of topics for the three corpus

compares the results with and without parameter reestimation and, therefore, with asymmetric and symmetric priors, $i = 10$ and $i = 0$, respectively.

Results in Fig. 5 show that when allowing asymmetric priors the variability and RGS metrics did not differ from those obtained using symmetric priors. The exact same results were obtained for the NIH and PATSTAT corpora. This fact allows us to conclude that, at least for the Mallet implementation and for reasonably large datasets, such as the ones used in our experiments, hyperparameter reestimation and asymmetric priors have a negligible effect on the resulting LDA models.

The last experiment measures the coherence of the topics, as this is one of the most widely used metrics for topic model evaluation. For this paper, C_V coherence was used, as it is the one that attains the highest correlation with all available human topic ranking data (Röder et al. 2015). Coherence values for the three real datasets are illustrated in Fig. 6. For the datasets of scientific papers and project proposals, coherence seems to increase or stabilize when increasing α and the number of topics. However, as previously discussed, a value of α that is too large has a negative effect on the RGS metric which is not captured by the coherence. Similarly, variability can actually increase if the number of topics is overestimated. When using coherence this is only noticed in the patents dataset, but in this case for the most likely values of α the number of topics that achieves the maximum coherence is below 25, which seems unrealistic given the richness of the patents dataset.

Our work shows that hyperparameter validation by means of graph metrics, both using variability and similarity to a reference graph, seems a promising way to optimize the design of topic models for STI text collections. Not only do these metrics provide richer information as compared with coherence, but they are also better aligned with some of the necessary goals for a wider adoption of these methodologies in real applications.

Conclusions and future lines

In this paper, we have proposed a novel methodology for the selection of topic models for large collections of documents, which is based on the analysis of the document similarities emanated from the topic models. We have shown that topic model selection criteria based on single metrics like the topic coherence may miss important aspects of the modeling process, like the variability of the document similarities under different runs of the modelling algorithm, or the quality of the similarity metric with respect to a reference graph. In this regard, two metrics have been proposed to address these aspects.

By analyzing the performance of the proposed methodology on a synthetic dataset and three real-world corpora belonging to the STI domain, we concluded that the proposed metrics provide insights about model variability and quality that can be exploited for hyperparameter selection. When compared with other schemes for the validation of topic models in this field, the advantage of our proposal is that it focuses precisely on the desirable features that actually matter to the the end users, rather than on theoretical details of the underlying LDA technique. As a consequence, our proposal can provide an important step forward in the adoption of these systems. In the end, the final choice might depend on the requirements of the specific application at hand, and could be affected by further subjective evaluation of the topics. However, we are convinced that a selection based on graph analysis tools provides a multidimensional view on the quality of the topic models which is useful to uncover information that might stay hidden from methods based on a single criterion such as coherence.

Three lines of further research can be outlined: (1) explore refinements of the variability metrics that take into account changes in the distributions of the document similarity values due to denser or sparser document topic-vectors. (2) the extension of the proposed methodology to the validation of dynamic topic models, and (3) the subjective evaluation by domain experts of the topic models emanated from the new metrics. This evaluation in different specific domains is planned as part of the European Union-funded project IntelComp[7].

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004870 (IntelComp project), and has also been partially supported by FEDER/ Spanish Ministry of Science, Innovation and Universities, State Agency of Research, project TEC2017-83838-R.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

[7] https://intelcomp.eu.

References

Adebiyi, A., Ogunleye, O. M., Adebiyi, M., & Okesola, J. (2019). A comparative analysis of tf-idf, lsi and lda in semantic information retrieval approach for paper-reviewer assignment. *Journal of Engineering and Applied Sciences, 14*(10), 3378–3382.

Agerri, R., Bermudez, J., & Rigau, G. (2014). Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC, 2014*, 3823–3828.

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology, 98*, 74–88.

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H. H., Peters, M., Power, J., Skjonsberg, S., Wang, L. L., Wilhelm, C., Yuan, Z., van Zuylen, & M., Etzioni, O. (2018) Construction of the literature graph in semantic scholar. In *NAACL*

Badenes-Olmedo, C., Redondo-Garcia, J. L., & Corcho, O. (2017). Distributing text mining tasks with librAIry. In *Proceedings of the 2017 ACM symposium on document engineering, DocEng '17* (pp. 63–66). ACM.

Badenes-Olmedo, C., Redondo-García, J. L., & Corcho O. (2020). Large-scale semantic exploration of scientific literature using topic-based hashing algorithms. *Semantic Web*, *11*, 735–750.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Procesing Systems, 18*, 147.

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

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., & Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE, 6*(3), e18029.

Burghardt, M., & Luhmann, J. (2021) Same same, but different? On the relation of information science and the digital humanities a scientometric comparison of academic journals using lda and hierarchical clustering

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M.(2009) Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Chen, J., Chen, J., Zhao, S., Zhang, Y., & Tang, J. (2020). Exploiting word embedding for heterogeneous topic model towards patent recommendation. *Scientometrics, 125*(3), 2091–2108.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J.(2015) Topiccheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 175–184).

Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T., & Waltman, L. (2021). A scientometric overview of cord-19. *PLoS ONE, 16*(1), e0244839.

European-Patent-Office: Data catalog patstat global. (2020). Data retrieved from the European Patent Office, https://www.epo.org/.

Grant, J., Hinrichs, S., Gill, A., & Adams, J.(2017) The nature, scale and beneficiaries of research impact

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate lda models? *Information Processing & Management, 54*(6), 1292–1307.

Han, X. (2020). Evolution of research topics in lis between 1996 and 2019: An analysis based on latent dirichlet allocation topic model. *Scientometrics, 125*(3), 2561–2595.

Hecking, T., & Leydesdorff, L. (2019). Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared with semantic maps. *Research Evaluation, 28*(3), 263–272.

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X, 5*(1), 011007.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2–3), 93–118.

Mantyla, M. V., Claes, M., & Farooq, U.(2018) Measuring LDA topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 1–4)

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet. cs. umass. edu

Miyata, Y., Ishita, E., Yang, F., Yamamoto, M., Iwase, A., & Kurata, K. (2020). Knowledge structure transition in library and information science: Topic modeling and visualization. *Scientometrics, 125*(1), 665–687.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T.(2010) Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108). Association for Computational Linguistics, Los Angeles, California

OECD: The Digitalisation of Science and Innovation Policy. (2018). https://www.oecd-ilibrary.org/content/component/sti_in_outlook-2018-17-en

Pathik, N., Shukla, P.(2020) Simulated annealing based algorithm for tuning LDA hyper parameters. In *Soft Computing: Theories and Applications* (pp. 515–521). Springer

Pérez-Fernández, D., Arenas-García, J., Samy, D., Padilla-Soler, A., & Gómez-Verdejo, V. (2019). Corpus viewer: NLP and ML-based platform for publicpolicy making and implementation.

Ranaei, S., Suominen, A., Porter, A., & Carley, S. (2020). Evaluating technological emergence using text analytics: two case technologies and three approaches. *Scientometrics, 122*(1), 215–247.

Röder, M., Both, A., & Hinneburg, A. (2015) Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408).

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467.

Srivastava, A., & Sutton, C. (2017) Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology, 67*(10), 2464–2476 (**Project code: 101488**).

Syed, S., Spruit, M.(2017) Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 165–174. IEEE

Vega-Carrasco, M., O'sullivan, J., Prior, R., Manolopoulou, I., & Musolesi, M.(2020) Modelling grocery retail topic distributions: Evaluation, interpretability and stability. arXiv preprint arXiv:2005.10125

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies, 1*(2), 691–713.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., et al. (2005). Database resources of the national center for biotechnology information. *Nucleic Acids Research, D33*(Database Issue), 39–D45.

Xiao, H., Stibor, T.(2010) Efficient collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of 2nd Asian Conference on Machine Learning, pp. 63–78. JMLR Workshop and Conference Proceedings*.

Xue, M.(2019) A text retrieval algorithm based on the hybrid lda and word2vec model. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)* (pp. 373–376). IEEE

Yao, L., Mimno, D., McCallum, A. (2009)Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09* (pp. 937–946). Association for Computing Machinery, New York, NY, USA

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. In *BMC bioinformatics* (Vol. 16, pp. 1–10). Springer