Central European Journal of Economic Modelling and Econometrics

Analysing the Impact of Removing Infrequent Terms on Topic Quality in Latent Dirichlet Allocation Models

Victor Bystrov[*], Viktoriia Naboka-Krell[†],
Anna Staszewska-Bystrova[‡], Peter Winker[§]

Submitted: 4.12.2024, Accepted: 27.05.2025

Abstract

An initial procedure in text-as-data applications is text preprocessing. One of the typical steps, which can substantially facilitate computations, consists in removing infrequent terms believed to provide limited information about the corpus. Despite the popularity of vocabulary pruning, there are not many guidelines on how to implement it in the literature. The aim of the paper is to fill this gap by examining the effects of removing infrequent terms for the quality of topics estimated using latent Dirichlet allocation (LDA). The analysis is based on Monte Carlo experiments taking into account different criteria for infrequent term removal and various evaluation metrics. The results indicate that pruning is often beneficial and that the share of vocabulary that might be eliminated can be quite considerable.

Keywords: topic models, text analysis, latent Dirichlet allocation, Monte Carlo simulation, text generation, text preprocessing

JEL Classification: C49

[*] University of Lodz, Poland; e-mail: victor.bystrov@uni.lodz.pl; ORCID: 0000-0003-0980-2790
[†] Justus Liebig University Giessen, Germany; e-mail: viktoriia.naboka@wirtschaft.uni-giessen.de; ORCID: 0000-0003-0690-2737
[‡] University of Lodz, Poland; e-mail: anna.bystrova@uni.lodz.pl; ORCID: 0000-0002-3941-4986
[§] Justus Liebig University Giessen, Germany; e-mail: peter.winker@wirtschaft.uni-giessen.de; ORCID: 0000-0003-3412-4207

DOI: 10.24425/cejeme.2025.155564

1 Introduction

The use of topic modelling techniques, especially latent Dirichlet allocation (LDA) introduced by Blei et al. (2003), is growing fast. The methods enable the analysis of large collections of texts in an unsupervised manner by uncovering latent structures (topics) behind the data. They find application in a broad variety of domains, including economics and econometrics (see e.g. Edison and Carcel, 2021; Bystrov et al., 2024b; Adämmer et al., 2025).

Given this increasing use of LDA as a standard tool for empirical analysis, the interest in details of the method, and, in particular, in parameter settings for its implementation is also rising. Thus, since the introduction of the LDA approach, different methodological components have been studied in more detail as, for example, the choice of the number of topics (Cao et al., 2009; Mimno et al., 2011; Lewis and Grossetti, 2022; Bystrov et al., 2024a), hyper-parameter settings (Wallach et al., 2009), model design (e.g. hierarchical structure as proposed by Teh et al., 2006), and inference methods (Griffiths and Steyvers, 2004).

However, not only the setting of technical parameters of the LDA model and the estimation algorithms are crucial for the results obtained, e.g. the identified topics. As the algorithm behind LDA "learns" from data based on co-occurrences of terms within texts, these have to be prepared in an appropriate way. LDA requires the text data to be structured in a document-term matrix (DTM), where each row corresponds to a document and each column to a specific term used throughout all documents. Then, the entry in a cell of the matrix provides the frequency of the term within a certain document. To obtain this matrix, the documents in a text corpus are usually cleaned and each document is represented as a bag-of-words (BoW), i.e. the algorithm neglects the semantic relationships between words and sentences. Together, these steps are referred to as *preprocessing*. By removing irrelevant terms and merging very similar terms (e.g. singular and plural forms of the same noun), preprocessing helps to reduce both the dimension of the DTM and its sparsity, which affect the performance of the algorithms used to estimate the LDA model (Maier et al., 2018, 2020).

Even though text preprocessing is an inherent component of any LDA analysis, there appear to be no common standards on how to perform it. In their illustrative application, Blei et al. (2003), for example, mention removing a standard list of stop words and all terms with an absolute frequency of one, i.e. showing up only once in the full corpus. In fact, such a step is usually performed in the majority of text-as-data applications with different lists of stop words and alternative rules for removing low- and – sometimes also – high-frequency terms. However, only a few attempts have been made so far to analyze the impact of text preprocessing on uncovered topics.

Denny and Spirling (2018) address this question by examining the impact of different combinations of text preprocessing steps on the outcomes of unsupervised techniques, including LDA. They show sensitivity of the results to preprocessing decisions;

however, since the analysis is done using real datasets, they cannot draw more general conclusions. The authors highlight the importance of careful data preparation for unsupervised techniques, like LDA, because, unlike for supervised methods, the results cannot be evaluated in a well-defined procedure (e.g. through accuracy measures as in text classification tasks). Lu et al. (2017) focus on removing terms that occur only once and those that are frequently used in three different real-world datasets (medical abstracts, articles published in biomedical journals, bibliographic records and abstracts from Elsevier Arts & Humanities journals). They measure the impact of removing (in)frequent terms by means of four different metrics. All in all, the authors come to the conclusion that removing singly occurring terms (i.e. the reduction of the vocabulary size by 30% to 40% depending on the underlying dataset) does not impact the topic modelling outcome substantially. Schofield et al. (2017) conduct some experiments to test the effect of removing common terms on topic quality using two datasets, the United States State of the Union Addresses and the annotated corpus of the New York Times. The authors conclude that removing stop words prior to model estimation does not impact topic inference.

Tang et al. (2014) analyze the properties of the data that affect the inferential performance of LDA models. They conduct small-scale Monte Carlo experiments using an LDA generative process with varying parameter configurations. In each experiment Tang et al. (2014) generate 30 corpora and compare true and estimated topics. Although they do not study the effects of text preprocessing, the results of their analysis elucidate the deterioration effect of data sparsity on the performance of LDA models.

A growing number of studies examine the consequences of text preprocessing on the results of supervised techniques (see e.g. Alam and Yao, 2019; Barushka and Hajek, 2020; Reimann and Dakota, 2021; HaCohen-Kerner et al., 2020; Al Sharou et al., 2021). These studies show that preprocessing can improve the performance of machine learning classifiers. They also highlight that each preprocessing procedure and each combination of preprocessing steps may matter for the final results and indicate the need for further systematic studies of initial text preparation.

In this contribution, we focus on the impact of removing terms with low frequency on the results of LDA modelling. Usually, low-frequency terms make up a large proportion of unique terms occurring in a corpus. This feature common to many, if not all languages can be approximated by Zipf's law, stating in its simplest version that term frequency is proportional to the inverse of the term frequency rank. A slightly more complex model has been proposed and estimated by Mandelbrot (1953). However, terms occurring only with low frequency are believed to be too specific to contribute to the meaning of the resulting topics when applying the LDA algorithm. Additionally, removing those terms substantially decreases the vocabulary size and, consequently, accelerates model estimation.

To the best of our knowledge, little research has been done so far to analyze the impact of removing infrequent terms on LDA estimation results. Maier et al. (2020)

focus on the consequences for topic quality of removing both frequent and infrequent words. They conduct their experiments on three different real-world datasets and conclude that vocabulary pruning does not qualitatively impact the resulting topics. To contribute to this line of research, we conduct a Monte Carlo simulation study. First, we define the characteristics of the data generating processes (DGPs) and following the generative model described by Blei et al. (2003) create true document-topic and topic-term distributions. For each of the DGPs, we generate a total of 1 000 pseudo-corpora. Finally, we apply different techniques, which have been proposed in the literature, to define and remove infrequent terms. Afterwards, LDA models are estimated based on the preprocessed corpora. These fitted models are misspecified as some parameters of the DGP are omitted. This leads to estimation bias, however at the same time, removal of noisy data reduces the variance of the estimator. Our simulations provide information on how these two effects as well as different settings for removing infrequent terms impact the estimation results.

The remainder of this paper is structured as follows. Section 2 introduces the steps that are usually performed for text data under the heading of text preprocessing. Focusing on removing infrequent terms, Section 3 describes the design of our Monte Carlo study. Next, in Section 4, we present and discuss the results of the experiments. Section 5 concludes.

2 Preprocessing of text data

Since texts are considered a very unstructured data source, text preprocessing usually precedes all other steps in text-as-data applications, regardless of the field of use. In general, these preprocessing steps can be divided into standard preprocessing steps and corpus or domain-specific preprocessing steps. In our description of changes applied to the vocabulary, we refer to “terms” as unique tokens included in the vocabulary and “words” as non-unique tokens in documents.

The standard preprocessing steps include the following: removing punctuation, special characters, and numbers; lowercasing; removing language specific stop words; lemmatizing or stemming. This list can be adjusted or extended by the so-called domain-specific preprocessing steps. For example, the character “#” falls into the category of special characters, but keeping it can be useful when working with Twitter data. In addition, the removal of extremely frequent and rare terms (relative pruning) could facilitate topic modeling.

Very frequent terms, also called corpus-specific stop words, occur in the majority of all documents and are often considered to be insufficiently specific to be useful for topic identification. Therefore, Grimmer and Stewart (2013) and Maier et al. (2018) remove all terms that appear in more than 99% of all documents.

Denny and Spirling (2018) provide two rationales for removing very rare terms: First, these terms contribute little information for topics retrieval, and, second, their removal reduces the size of the vocabulary substantially and, consequently, speeds up

computations. A common rule of thumb, mentioned in Denny and Spirling (2018), is to discard terms that appear in less than 0.5-1% of the documents. Denny and Spirling (2018) notice, however, that there has been no systematic study of the effects this preprocessing choice has on the modeling of the topics.
Infrequent terms can be removed using one of the following criteria:

i) Document frequency: remove terms with the frequency of showing up across the documents in the corpus below the defined threshold (absolute/relative).

ii) Term frequency: remove terms with frequency in the corpus below the defined threshold (absolute/relative).

iii) Term frequency-inverse document frequency (TF-IDF) values describing relative importance of terms for specific documents: remove terms with low TF-IDF values (Blei and Lafferty, 2009).

There are no obvious rules for setting the required thresholds. Grimmer and Stewart (2013) notice that the choice of thresholds for removing common and rare terms from a corpus should be contingent on the diversity of the vocabulary, the average length of documents and the size of the corpus. However, this is a heuristic observation that is not based on a systematic analysis.

3 Monte Carlo study design

To analyze the impact of removing infrequent terms in the context of LDA in a systematic way, we conduct a Monte Carlo simulation study. The purpose of the analysis is to provide insight into the effects of vocabulary pruning on topic quality in estimated LDA models. Given that the actual topics are known in the experiments, we focus in particular on the difference between the estimated and true topics. Obviously, this difference is driven only to some extent by the specific preprocessing used, but depends also on the sampling error, which we have to take into account when summarizing our findings.
In this section, we first describe the setup of simulation experiments. Then, we present the features of the DGPs and details of the procedure of corpora generation (subsection 3.1). Afterwards, we define and describe the rules for the removal of infrequent terms to be applied in the Monte Carlo study (subsection 3.2). Finally, we discuss different quality measures used to evaluate the results (subsection 3.3). Code details for data generation, model estimation, and evaluation is available on Github at `https://github.com/VikaNa/removing-infrequent-words-lda`.

3.1 Corpora generation

We start by presenting two DGPs to be considered in the Monte Carlo study. The choice of settings for DGPs is a trade-off between using data dimensions that are often encountered in economic applications and computational costs of Monte Carlo simulations. Textual data (policy reports, research articles, news, social media etc.) are instrumental in measuring economic sentiments and expectations as well as forecasting indicators of actual economic and financial activity. The features of a selected textual corpus depend on its relevance for an econometric application. We consider two types of DGP that mimic some characteristics of text corpora encountered in economic applications, but limit the data dimensions to accommodate the high computational costs of Monte Carlo simulations.

Table 1 summarizes the main characteristics of selected DGPs. The first one contains a relatively small number of long documents covering a moderately large number of topics. These characteristics are derived from some real-world datasets such as policy reports and research articles which are used for evaluating policy communications and making long-term economic projections (e.g. Hansen et al., 2018; Hartmann and Smets, 2018). DGP2 has the characteristics of corpora containing a large number of short texts discussing a relatively small number of topics. They are typical for collections of news articles and microblog posts which are often used for measuring expectations and sentiments as well as the short-term forecasting of economic and financial indicators (e.g. Lüdering and Tillmann, 2020; Angelico et al., 2022).

Table 1: Characteristics of DGPs

	#documents	# words per document	# unique terms	# topics, K
DGP1	1,000	3,000	30,000	50
DGP2	10,000	150	20,000	15

Given these features of the DGPs, we follow the generative probabilistic model described by Blei et al. (2003). The underlying assumption of the model is that each text corpus contains latent topics that can be represented by probability distributions of vocabulary terms. Each document in a corpus can be represented by a mixture of topic-term probability distributions that realizes the document-term distribution.

For each DGP, the matrix of topic-term probabilities $\boldsymbol{\beta}$ is drawn from the Dirichlet distribution using a single concentration parameter $\eta = 1/K$. Algorithm 1 describes how each document $\mathbf{w}$ in a corpus D is generated. The length of the document N is defined by drawing from a Poisson distribution where the parameter ξ is equal to the expected number of words in a document, namely 3,000 for DGP1 and 150 for DGP2. For each document $\mathbf{w}$, the vector of topic probabilities θ is drawn from the Dirichlet distribution using a concentration parameter $\alpha = 1/K$. For each word in a document, a topic z_n is first drawn from the multinomial distribution parametrized by vector θ and then a term w_n is drawn from the multinomial distribution given the topic z_n and the matrix of topic-term probabilities $\boldsymbol{\beta}$. The choice of flat priors $1/K$ for the parameters α and η is in accordance with the default parameter setting in software implementations, e.g. in Python's `scikit-learn` library (Pedregosa et al., 2011). These default values are used in many text-as-data applications.

Algorithm 1 Generative probabilistic model by Blei et al. (2003)

```
Choose $\boldsymbol{\beta} \sim Dir(\eta)$
  for document $\mathbf{w}$ in corpus $D$ do
    Choose $N \sim Poisson(\xi)$
    Choose $\theta \sim Dir(\alpha)$
    for word $w_n = 1, 2, \ldots, N$ do
      (a) Choose a topic $z_n \sim Multinomial(\theta)$
      (b) Choose a term $w_n$ from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial
            probability distribution conditioned on the topic $z_n$
    end for
  end for
```

We use Algorithm 1 to generate 1 000 different pseudo-corpora for each DGP. While the characteristics of the selected DGPs are aligned with real corpora in economics, the generated synthetic corpora in our experiment cannot be further described and interpreted. The obtained pseudo-corpora do not contain real words, but synthetic words such as "word1", "word256" etc.

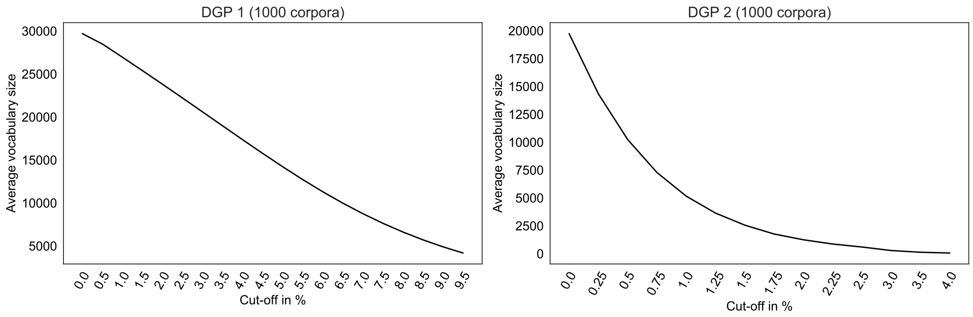
3.2 Removal of infrequent terms

A popular approach to vocabulary pruning is to remove all terms that appear in a small number of documents in the corpus. As indicated in Section 2, this criterion can be based on the absolute number of documents (e.g., remove all terms that occur in no more than one document) or the relative number of documents (e.g., remove all terms that occur in no more than in 1 percent of all documents in the corpus). In the Monte Carlo experiments we consider different values of the relative cut-off for removing terms on the basis of relative document frequency. We use Python's `scikit-learn` (version 0.24.2) library and its *CountVectorizer()* class to exectute this task.

Before fixing the range of cut-off values, we consider the resulting distribution of the vocabulary size for each DGP: Figure 1 shows the average vocabulary size over 1 000 corpora as a function of the relative cut-off value (relative document based frequency) for each DGP. For the cut-off value of 1% that is often used in empirical applications, the vocabulary size decreases by 9.3% and 74.1% for DGP1 and DGP2, respectively. Given these differences in the relative distributions of vocabulary sizes for selected

DGPs, in the simulations, we use different ranges of cut-off values. For DGP1, we proceed in steps of 0.5% within the interval [0.0%; 9.5%]. For DGP2, we reduce the step size to 0.25% up to the cut-off value of 2.5% and set the maximum cut-off value to 4% because higher thresholds would result in an empty vocabulary.

Figure 1: Average vocabulary size depending on the relative cut-off value



For every corpus generated from DGP1 we build 20 different sub-samples, according to the defined cut-off values. We use each subsample to estimate an LDA model. For every corpus obtained from DGP2, 14 different sub-samples are constructed and corresponding LDA models are estimated. For model estimation, we use Python's library `scikit-learn` (version 0.24.2). We leave all parameters at their default values except the number of topics, which is set according to the DGP characteristics defined in Table 1 and the maximum number of iterations, which is set to 100 due to computational constraints. In addition, we fix a random state for the reproducibility of the results.

As described in Section 2, there are two additional criteria for vocabulary pruning that are frequently used in applications: term frequency and TF-IDF frequency. We use them to perform robustness checks for the results presented in our study. The results of this robustness analysis are presented in Appendix A.

3.3 Evaluation

Throughout each Monte Carlo scenario, we keep all the parameters constant, except for the document-term matrix required as input for the estimation of the LDA model. As described in the previous subsection, different variations of one corpus are created by applying various cut-off values to remove infrequent terms. As a result, we obtain 20 and 14 LDA models for DGP1 and DGP2, respectively.

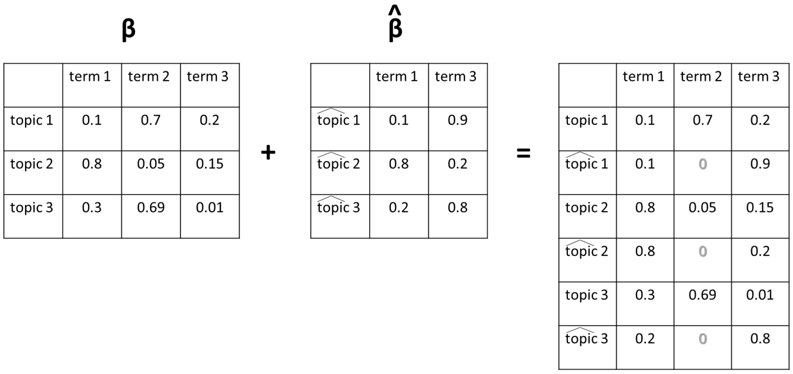
Different evaluation techniques have been developed to access topic modelling quality. Some of them became standard in different text-as-data applications, e.g. topic coherence (Mimno et al., 2011) or topic similarity (Cao et al., 2009). The measure of Mimno et al. (2011) was designed to correspond with the judgement of the consistency of topics by humans. It is based on maximization of the average semantic coherence across a range of topics. The method of Cao et al. (2009) associates good topic quality with sharp topic distinction or lack of overlap. The proposed measure is computed by minimizing the average cosine similarity between each pair of topics. Another popular measure used to evaluate the model's predictive performance on an unseen (or held-out) sample is perplexity. It is defined as the inverse of the geometric mean per-word likelihood. Blei et al. (2003) show that perplexity is monotonically decreasing in the likelihood of the test data with increasing number of topics. Reducing the size of the vocabulary while keeping the number of topics constant leads qualitatively to the same effects. For this reason, in the current study, we do not consider perplexity as an evaluation metric.

Instead, we also compute recall (or the share of reproduced topics) as proposed by by Bystrov et al. (2024a) and model fit to evaluate the impact of removing infrequent terms on topic quality in LDA models.

First, using the *recall* metric, we aim to measure how the true structure of topics changes (by comparing *true* and *estimated* topic-term distributions). In the current work, we follow a similar approach to the one proposed by Bystrov et al. (2022) and apply the so-called *best matching*:

1. Compare true and estimated topic-term distributions based on the union of two vocabularies. For terms not contained in the estimated topic-term distribution, assign probability of zero. An example of this procedure is presented in Figure 2.

Figure 2: Best Matching: example



β

	term 1	term 2	term 3
topic 1	0.1	0.7	0.2
topic 2	0.8	0.05	0.15
topic 3	0.3	0.69	0.01

+

$\hat{\beta}$

	term 1	term 3
$\widehat{\text{topic 1}}$	0.1	0.9
$\widehat{\text{topic 2}}$	0.8	0.2
$\widehat{\text{topic 3}}$	0.2	0.8

=

	term 1	term 2	term 3
topic 1	0.1	0.7	0.2
$\widehat{\text{topic 1}}$	0.1	0	0.9
topic 2	0.8	0.05	0.15
$\widehat{\text{topic 2}}$	0.8	0	0.2
topic 3	0.3	0.69	0.01
$\widehat{\text{topic 3}}$	0.2	0	0.8

2. For each of the estimated topics, calculate *similarity/distance* to each of the true topics. Then assign the true topic with the highest (lowest) similarity (distance).

3. Define and apply a cut-off value to keep good quality matches only. Calculate the *recall* metric as the share of correctly reproduced topics.

In their empirical application, Bystrov et al. (2022) use cosine similarity in step 2 and automatically determine a data-based cut-off as the 95% percentile of all pairwise similarity scores in step 3. Maier et al. (2020), who also studied the impact of removing infrequent terms on topic quality, perform topic matching based on top 20 topic terms following the approach proposed by Niekler and Jähnichen (2012). The authors calculate pairwise cosine distances and apply a cut-off value of 0.5 to obtain the share of reproduced topics.

In the current application, we use different metrics to measure the similarity between true and estimated topics:

a) cosine similarity: takes values between -1 (two vectors point in opposite directions) and 1 (two vectors point in the same direction).

b) Jensen-Shannon divergence/distance: ranges between 0 (two distributions are the same) and 1 (two distributions are completely different).

c) rank-biased overlap (RBO) proposed by Webber et al. (2010) to compare ranked lists: ranges from 0 (ranked lists are disjoint) to 1 (ranked lists are exactly the same).

Since the true topics appear to be very distinct from each other in the current Monte Carlo study, we decided to use a cut-off value of 0.8 for the similarity metrics (cosine similarity and RBO) and 0.2 for the distance metric (Jensen-Shannon).

Alternatively, one can use *one-to-one matching* as described by Bystrov et al. (2022). The resulting measure is called *model fit*. Thereby, all of the topics have to be matched using the Hungarian algorithm and a defined distance metric. Matches are assigned to minimize the overall cost of assignment. Thus, the mean distances between the identified matches can be considered to measure the quality of the fit of the model. In the case of simulated data, the model fit metric based on one-to-one matching offers a different focus on the defined problem. Since the true number of topics is known, it is of special interest to see how the true and estimated topics are assigned to each other when none of the topics is left out.

4 Results

In this section, we summarize the main findings of the Monte Carlo analysis. Thereby, we focus on the removal of infrequent terms according to their document frequency in the corpus as described in Section 3.2. The results presented here are based on 1 000 replications. We also perform robustness checks using 100 replications for the alternative criteria for vocabulary pruning, namely absolute term frequency and TD-IDF values, and present the results in Appendix A.

Figures 3 and 4 present the metrics values obtained after document frequency pruning. The cutoff values exhibited on the x axis in Figures 3 and 4 describe the minimum share of documents in which a term must be included to not be removed from the corpus. Thus, a cut-off value of 0.0% corresponds to keeping all terms (30K for DGP1 and almost 20K for DGP2), while 9.5% in Figure 3 refers to the removal of all terms which do not show up in at least 9.5% of all documents leaving only about 4K terms in the corpus. Consequently, in Figure 4, the value of 4.0% corresponds to keeping only those terms, which appear in at least 4.0% of all documents reducing the size of the vocabulary to 60 terms.

On the ordinate, Figures 3 and 4 show, as solid lines, the means of the evaluation metrics obtained over 1 000 replications for DGP1 and DGP2, respectively. The dashed lines in the first three subplots provide the 20% and 80% quantiles of the distributions of these metrics. The corresponding bands for the measures from the last panel (*recall*) are shown in Figures A7 and A8 in Appendix B. The metrics considered include: *model fit* (Bystrov et al., 2024a) (to be minimized), topic similarity (Cao et al., 2009) (to be minimized), topic coherence (Mimno et al., 2011) (to be maximized), and *recall* (to be maximized). In empirical applications, the true DGPs and corresponding topics are unknown. Thus, the *recall* criteria cannot be applied. The observed collapse of *recall* for higher cut-off values indicates that the remaining vocabulary is no longer sufficient to identify the true topics.

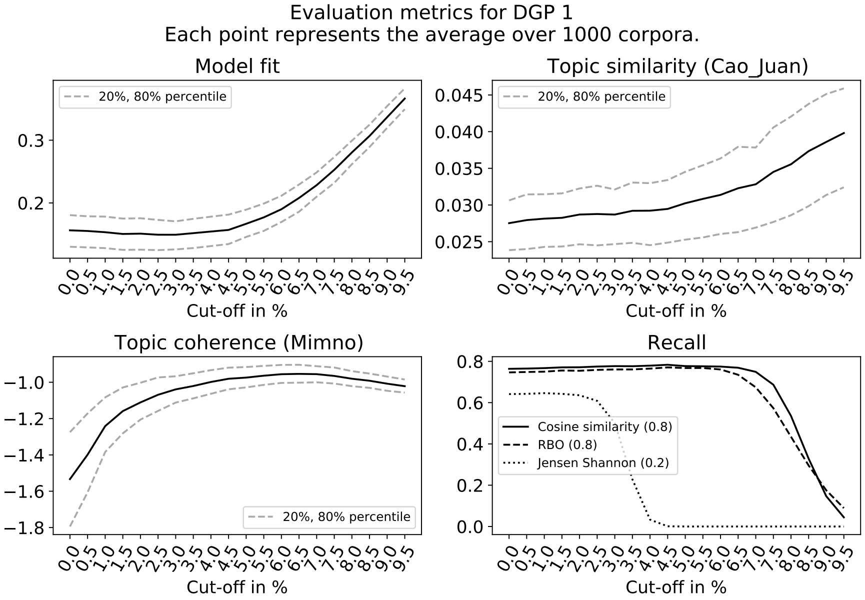
It becomes obvious from Figures 3 and 4 that removing infrequent terms has consequences for the results of the LDA estimation. As a general pattern, we conclude that the application of pruning is beneficial for low cut-off values. This might be attributed to two effects. First, terms appearing only in a few documents do not contain much information about more general topics. Second, removing these terms substantially reduces the dimensionality of the estimation problem, which increases the efficiency of the estimators. However, beyond a certain point the increasing loss of information resulting from the removal of more and more infrequently used terms dominates the gains due to reduced dimensionality. Comparing the findings from Figures 3 and 4, it appears that gains and losses from decreasing vocabulary size by eliminating rare terms are weighted somewhat differently by alternative evaluation criteria.

For DGP1 (Figure 3), the lowest average distances between the true and estimated topic sets as measured by *model fit* correspond to cut-off values from 3% to 4.5%. Further removal of infrequent terms leads to increased distortions in estimated topics. The best values of *coherence* are obtained for thresholds of 3%-6.5%. The metric is quite sensitive to keeping too many infrequent terms in the texts, showing significantly smaller values for initial thresholds. In the case of *similarity*, thresholds up to 4.5% lead to similar metric values. Eventually, alternative versions of *recall* measure indicate that the maximum threshold that might be considered is about 3% (metric based on Jensen-Shannon distance) or 6.5% (cosine similarity and RBO-based metrics). Altogether, if all metrics are considered jointly, the best threshold is about

3%. A similar conclusion is reached if the TF-IDF or absolute term frequency-based vocabulary pruning is performed instead of the document frequency pruning (see Appendix A).

A similar analysis for DGP2 (Figure 4) suggests the following cut-off values. According to *model fit* the interval from 0.25% to 0.75% could be considered, while the *coherence* metric indicates the range 0.5%-2.5%. Topic *similarity* is quite similar for cut-off values up to 0.5% and *recall* metrics suggest stopping at 0.25%, 1.25% or 2% starting from the most restrictive measure. Thus, in general, a threshold of about 0.25%-0.5% might be selected. This finding is again quite robust with respect to the criterion used for the removal of infrequent terms (see Appendix A).

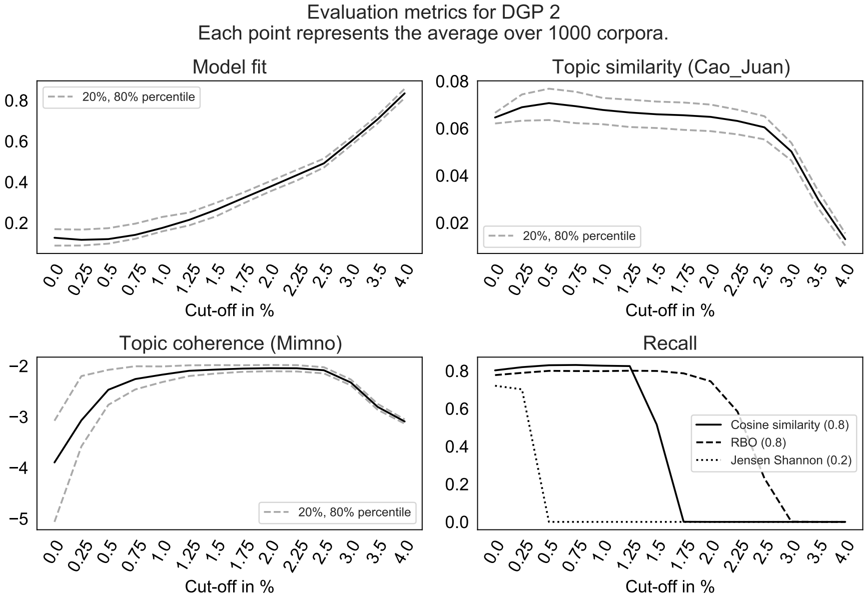
Figure 3: Evaluation of document frequency-based vocabulary pruning for DGP1



Notes: Model fit, topic similarity: lower values are preferred. Topic coherence, recall: higher values are preferred.

For a better understanding of the results from Figures 3 and 4, the selected thresholds were juxtaposed with the corresponding shares of terms removed from the vocabularies (see Figures A5 and A6 in Appendix B). The cut-off value of 3% for DGP1 corresponds to reducing the size of the vocabulary by 30% and cut-offs of 0.25-0.5% for DGP2 imply removing 27-48% of all terms. Thus, in both cases, it could be concluded that the reduction in vocabulary size, which could accelerate the estimation process without affecting qualitatively the results, was considerable and

Figure 4: Evaluation of document frequency-based vocabulary pruning for DGP2



amounted to approximately 30% of all terms. These results show that guidelines focusing on removing infrequent terms up to a certain share of all terms might be worth following up. We also compare our results to those obtained by Lu et al. (2017) who considered three different real-world datasets to analyze, among others, sensitivity of estimation results to removing singly occurring terms. The characteristics of our DGP1 are most similar to those of the *Genomics06* dataset (full-text articles published in biomedical journals), and the characteristics of DGP2 – to *Ohsumed* dataset (abstracts). Considering three different metrics (document space density, entropy, pairwise topic similarity), the authors conclude that removing singly occurring terms, constituting 41.56% of terms for *Genomics06* data and 30.77% for *Ohsumed* dataset, does not negatively impact the results. Removing only singly occurring terms, in our simulations, led to an average reduction in vocabulary size of 1% for DGP1 and 1.32% for DGP2. However, by applying the derived optimal cut-off values we obtained proportions of removed terms similar to the shares discussed by Lu et al. (2017).

5 Conclusions

The focus of this paper was on preprocessing of text data in the context of LDA analysis. Although text preprocessing is an essential part of data preparation in text-as-data applications and some rules-of-thumb of text preprocessing sequences exist and are often followed, there is only little evidence on how particular text preprocessing decisions might affect the final results. In the specific setting considered in this paper, the outcome of interest were the estimated topics and the analyzed preprocessing step was the removal of infrequent terms in a text corpus.

To allow for a systematic evaluation of the impact of different techniques for reducing vocabulary size and generalizable conclusions, we conducted a Monte Carlo simulation study. We first generated data from scratch based on two pre-defined DGPs following the probabilistic model proposed by Blei et al. (2003). For each of the defined DGPs, we then applied different techniques to remove rare terms from the texts and estimated multiple LDA models varying the text input only. Finally, we evaluated the results using some well established metrics such as *coherence* and *topic similarity* that focus on the estimated set of topics as well as *model fit* and *recall* metrics that are based on the comparison between the true and estimated set of topics.

Our results indicate that appropriate removal of infrequent terms can improve the LDA estimation results. This is caused by the reduction of dimensions and the sparsity of the document-term matrix paired with a limited loss of information about the content of the topics.

The results have at least two practical implications. First, we show that across the DGPs considered, about 30% of terms can be removed without qualitative losses in the resulting topics. This is a valuable insight for the scientists who work with substantial sets of data containing long texts on average. Most real-world data sets have large or even very large vocabularies. In such cases, removing 30% of terms could lead to a considerable decrease in computing time and an increase in efficiency. Second, we demonstrate the robustness of these conclusions with respect to the application of different techniques to reduce the size of vocabularies. This implies that in practice, vocabulary pruning can be based on either of the popular criteria.

Our results provide support for the common practice of vocabulary pruning. They shed light on the share of terms that might be removed from corpora meeting two criteria: being well described by an LDA model and having similar characteristics to our synthetic DGPs. It should be noted that we did not study collections of documents outside the LDA framework, i.e. such that would not be well approximated by an LDA model. Using our results for texts substantially different from those used in the simulations should also be done with caution.

The paper suggests that future research could follow the ideas of Denny and Spirling (2018) and focus on an evaluation of different combinations of text preprocessing steps as well as other DGPs. However, performing this analysis in a systematic manner by means of Monte Carlo experiments would require substantially more computational resources. For example, it might be worthwhile to consider the combined impact of stemming/lemmatizing and vocabulary pruning.

Acknowledgements

Financial support from the German Research Foundation (DFG) (WI 2024/8-1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO-2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

References

[1] Adämmer P., Prüser J., Schüssler R. A., (2025), Forecasting macroeconomic tail risk in real time: Do textual data add value?, *International Journal of Forecasting* 41(1), 307–320, available at `https://www.sciencedirect.com/science/article/pii/S0169207024000463`.

[2] Al Sharou K., Li Z., Specia L., (2021), Towards a better understanding of noise in natural language processing, [in:] *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, [eds.:] R. Mitkov and G. Angelova, INCOMA Ltd., Held Online, 53–62, available at `https://aclanthology.org/2021.ranlp-1.7`.

[3] Alam S., Yao N., (2019), The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis, *Computational and Mathematical Organization Theory* 25.

[4] Angelico C., Marcucci J., Miccoli M., Quarta F., (2022), Can we measure inflation expectations using twitter?, *Journal of Econometrics* 228(2), 259–277.

[5] Barushka A., Hajek P., (2020), The effect of text preprocessing strategies on detecting fake consumer reviews, *Proceedings of the 2019 3rd International Conference on E-Business and Internet, ICEBI '19*, Association for Computing Machinery, New York, NY, USA, 13–17, available at `https://doi.org/10.1145/3383902.3383908`.

[6] Blei D. M., Lafferty J. D., (2009), Topic models, [in:] *Text mining: classification, clustering, and applications*, [eds.:] AN. Srivastava, M. Sahami, CRC Press, Boca Raton, 71–94.

[7] Blei D. M., Ng A. Y., Jordan M. I., (2003), Latent Dirichlet allocation, *Journal of Machine Learning Research* 3, 993–1022.

[8] Bystrov V., Naboka-Krell V., Staszewska-Bystrova A., Winker P., (2024a), Choosing the number of topics in LDA models – A Monte Carlo comparison of selection criteria, *Journal of Machine Learning Research* 25(79), 1–30.

[9] Bystrov V., Naboka-Krell V., Staszewska-Bystrova A., Winker P., (2024b), Comparing Links between Topic Trends and Economic Indicators in the German and Polish Academic Literature, *Comparative Economic Research. Central and Eastern Europe* 2, 7–28, available at `https://czasopisma.uni.lodz.pl/CER/article/view/23389/2344417`.

[10] Bystrov V., Naboka V., Staszewska-Bystrova A., Winker P., (2022), Cross-corpora comparisons of topics and topic trends, *Journal of Economics and Statistics* 242(4), 433–469, available at `https://doi.org/10.1515/jbnst-2022-0024`.

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

[12] Denny M. J., Spirling A., (2018), Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it, *Political Analysis* 26(2), 168–189.

[13] Edison H., Carcel H., (2021), Text data analysis using latent Dirichlet allocation: an application to FOMC transcripts, *Applied Economics Letters* 28(1), 38–42, `https://doi.org/10.1080/13504851.2020.1730748`.

[14] Griffiths T. L., Steyvers M., (2004), Finding scientific topics, Proceedings of the National Academy of Sciences 101(suppl 1), 5228–5235.

[15] Grimmer J., Stewart B. M., (2013), Text as data: The promise and pitfalls of automatic content analysis methods for political texts, *Political Analysis* 21, 267–297.

[16] HaCohen-Kerner Y., Miller D., Yigal Y., (2020), The influence of preprocessing on text classification using a bag-of-words representation, *PLOS ONE* 15, e0232525.

[17] Hansen S., McMahon M., Prat A., (2018), Transparency and deliberation within the FOMC: A computational linguistics approach, *The Quarterly Journal of Economics* 133(2), 801–870.

[18] Hartmann P., Smets F., (2018), The european central bank's monetary policy during its first 20 years, *Brookings Papers in Economic Activity*, 1–118.

[19] Lewis C., Grossetti F., (2022), A statistical approach for optimal topic model identification, *Journal of Machine Learning Research* 23, 1–20.

[20] Lu K., Cai X., Ajiferuke I., Wolfram D., (2017), Vocabulary size and its effect on topic representation, *Information Processing & Management* 53(3), 653–665.

[21] Lüdering J., Tillmann P., (2020), Monetary policy on Twitter and asset prices: Evidence from computational text analysis, *The North American Journal of Economics and Finance* 51(C).

[22] Maier D., Niekler A., Wiedemann G., Stoltenberg D., (2020), How document sampling and vocabulary pruning affect the results of topic models, *Computational Communication Research* 2, 139–152.

[23] Maier D., Waldherr A., Miltner P., Wiedemann G., Niekler A., Keinert A., Pfetsch B., Heyer G., Reber U., Häussler T., Schmid-Petri H., Adam S., (2018), Applying LDA topic modeling in communication research: Toward a valid and reliable methodology, *Communication Methods and Measures* 12, 93–118.

[24] Mandelbrot B., (1953), An informational theory of the statistical structure of language, [in:] *Communication Theory*, [ed.:] W. Jackson, Academic Press, Princeton, 486–502.

[25] Mimno D., Wallach H., Talley E., Leenders M., McCallum A., (2011), Optimizing semantic coherence in topic models, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., 262–272, available at `https://aclanthology.org/D11-1024`.

[26] Niekler A., Jähnichen P., (2012), Matching results of latent Dirichlet allocation for text, [in:] *Proceedings of the 11th International Conference on Cognitive Modeling*, [eds.:] N. Rußwinkel, U. Drewitz, H. van Rijn, Universitaetsverlag der TU Berlin, 317–322.

[27] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay E., (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12, 2825–2830.

[28] Reimann S., Dakota D., (2021), Examining the effects of preprocessing on the detection of offensive language in German tweets, [in:] *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, [eds.:] K. Evang, L. Kallmeyer, R. Osswald, J. Waszczuk, T. Zesch, KONVENS 2021 Organizers, Düsseldorf, Germany, 159–169, available at `https://aclanthology.org/2021.konvens-1.14`.

[29] Schofield A., Magnusson M., Mimno D., (2017), Pulling out the stops: Rethinking stopword removal for topic models, *Proceedings of the 15th Conference of the*

European Chapter of the Association for Computational Linguistics, Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 432–436, available at `https://aclanthology.org/E17-2069`.

[30] Tang J., Meng Z., Nguyen X., Mei Q, Zhang M., (2014), Understanding the limiting factors of topic modeling via posterior contraction analysis, [in:] *Proceedings of the 31st International Conference on Machine Learning*, [eds.:] E. P. Xing, T. Jebara, vol. 32 of Proceedings of Machine Learning Research, PMLR, Bejing, China, 190–198, `https://proceedings.mlr.press/v32/tang14.html`.

[31] Teh Y. W., Jordan M. I., Beal M. J., Blei D. M., (2006), Hierarchical Dirichlet processes, *Journal of the American Statistical Association* 101(476), 1566–1581, available at `http://www.gatsby.ucl.ac.uk/ywteh/research/npbayes/jasa2006.pdf`.

[32] Wallach H. M., Mimno D., McCallum A., (2009), Rethinking LDA: Why priors matter, *Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09*, Curran Associates Inc., Red Hook, NY, USA, 1973–1981.

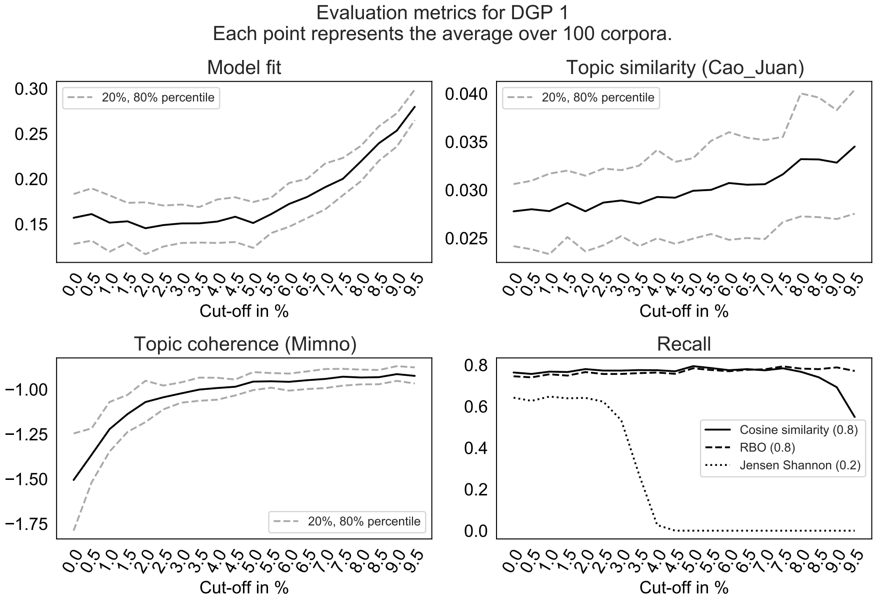
[33] Webber W., Moffat A., Zobel J., (2010), A similarity measure for indefinite rankings, *ACM Trans. Inf. Syst.* 28(4), available at `https://doi.org/10.1145/1852102.1852106`.

A Robustness checks

Term frequency

This approach to vocabulary pruning is based on the absolute frequency of terms in the considered corpus. The rule was applied e.g. by Blei et al. (2003) who removed all terms that occurred only once in the corpus used in their illustrative example. To make the results based on term frequency comparable to the results based on document frequency, we consider specific sequences of cut-off values for each DGP. These thresholds are such that the vocabulary sizes were comparable to vocabulary sizes implied by document frequency cut-off values. To compute them we identify vocabulary sizes corresponding to the relative cut-offs applied in document frequency based pruning. Then, we identify minimum absolute term frequencies corresponding to the considered relative cut-offs. Figures A1 and A2 show the results for DGP1 and DGP2, respectively.

Figure A1: Evaluation of absolute term frequency based vocabulary pruning for DGP1



Notes: Model fit, topic similarity: lower values are preferred. Topic coherence, recall: higher values are preferred.

TF-IDF

Blei and Lafferty (2009) propose to use TF-IDF to prune the vocabulary. In their experiments, they consider the top 10,000 terms with highest TF-IDF values. TF-IDF is a weighted measure that is used to determine the importance of a term for a given corpus and consists of two parts, namely term frequency (TF) and inverse document frequency (IDF):

$$\text{Term Frequency}_{w,D} = \frac{\text{Number of times term } w \text{ appears in document } D}{\text{Total number of term } w \text{ in document } D} \tag{1}$$

$$\text{Inverse Document Frequency}_{w} = log \frac{\text{Total number of documents}}{\text{Number of documents with term } w} \tag{2}$$

The IDF part accounts for terms that occur in the majority of documents (e.g. stop words) and scales down their importance. Finally, TF-IDF score is calculated by multiplying TF and IDF:

$$\text{TF-IDF}_{w,D} = \text{Term Frequency}_{w,D} * \text{Inverse Document Frequency}_{w} \tag{3}$$

For each of the corpora generated from DGP1, we build 20 different sub-samples considering the top V terms with the highest TF-IDF values. To make the results comparable, we choose V equal to the vocabulary size that results when document frequency-based rules are applied (see Figure 1). For example, if applying a document frequency cut-off value of 6 percent results in a vocabulary size of about 10,000 terms for corpus x, we consider only 10,000 terms with the highest TF-IDF values for this corpus. Figures A3 and A4 present the results based on TF-IDF vocabulary pruning.

Figure A2: Evaluation of absolute term frequency based vocabulary pruning for DGP2

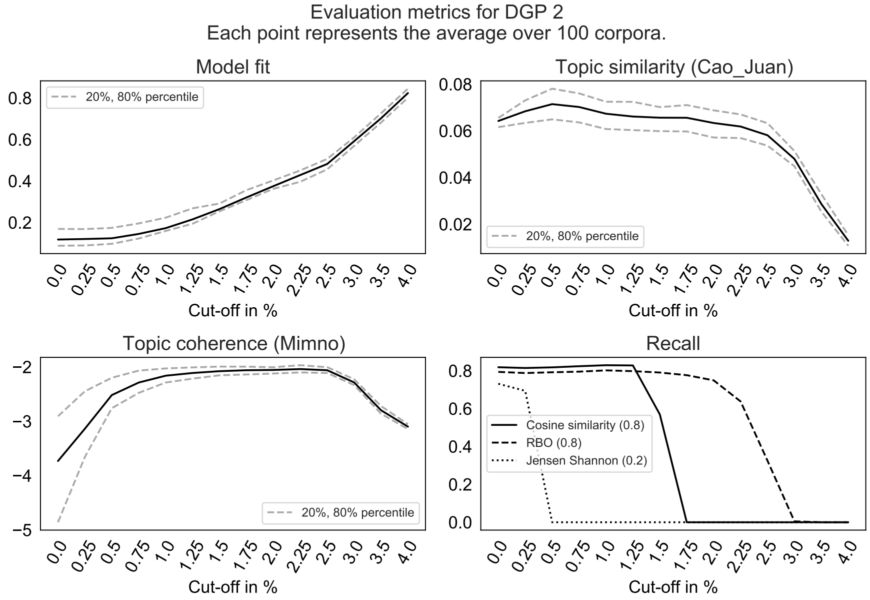


Figure A3: Evaluation of TF-IDF based vocabulary pruning for DGP1

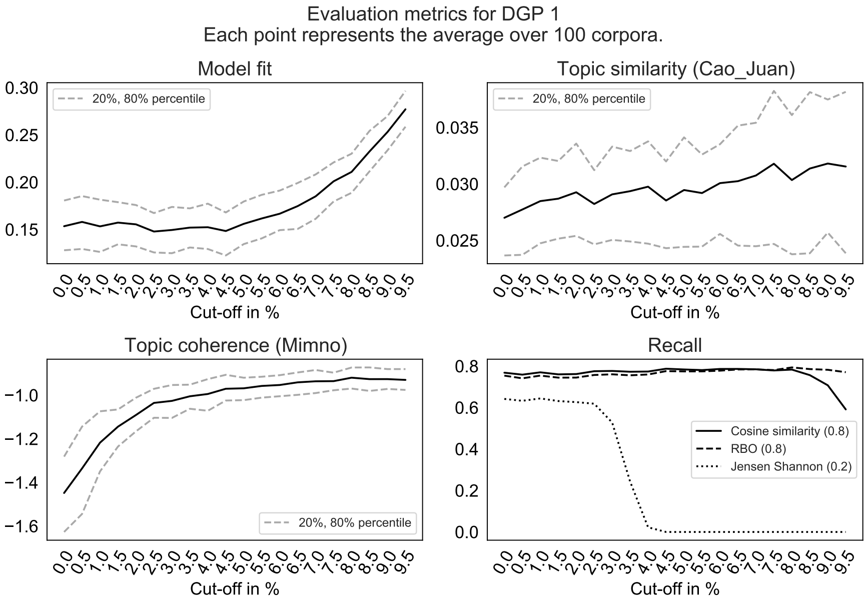
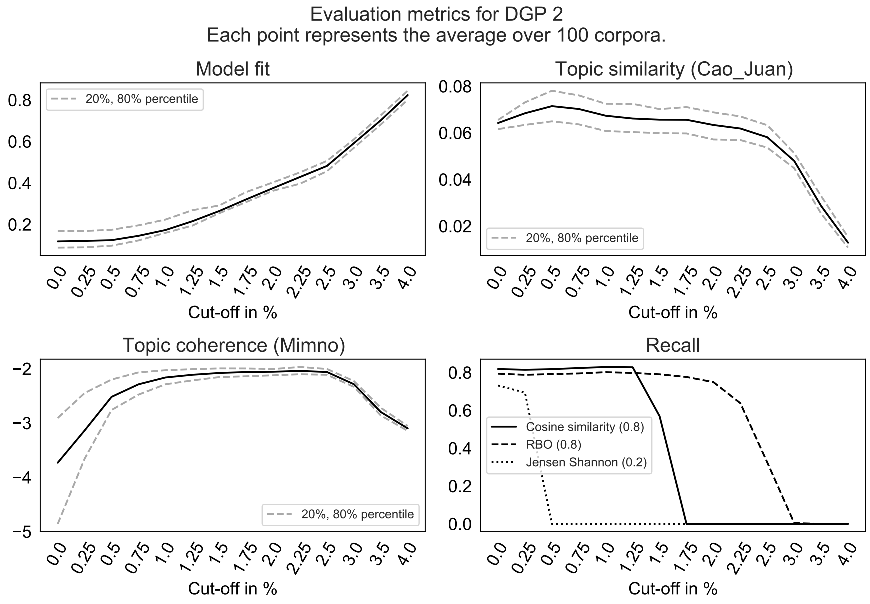


Figure A4: Evaluation of TF-IDF based vocabulary pruning for DGP2



B Additional visualizations

Figure A5: Recall values (higher values are preferred) and additional statistics for DGP1

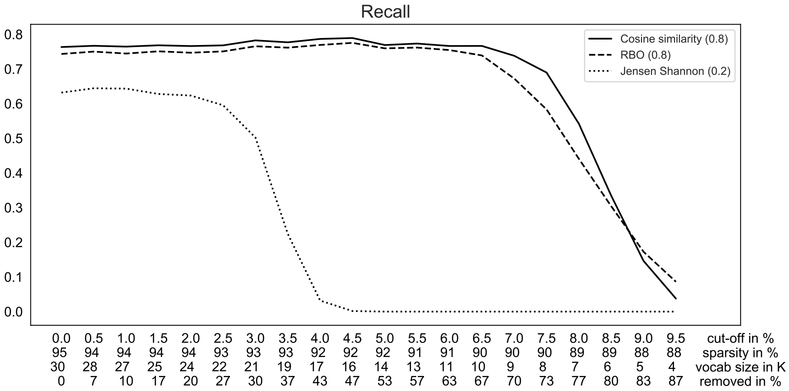


Figure A6: Recall values (higher values are preferred) and additional statistics for DGP2

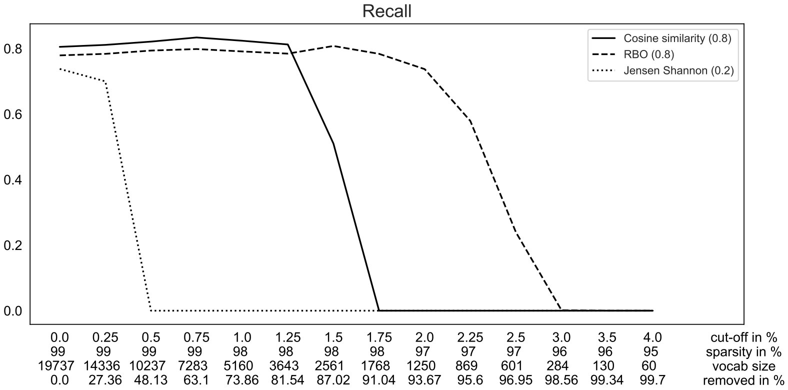


Figure A7: Recall values for DGP1 (higher values are preferred)

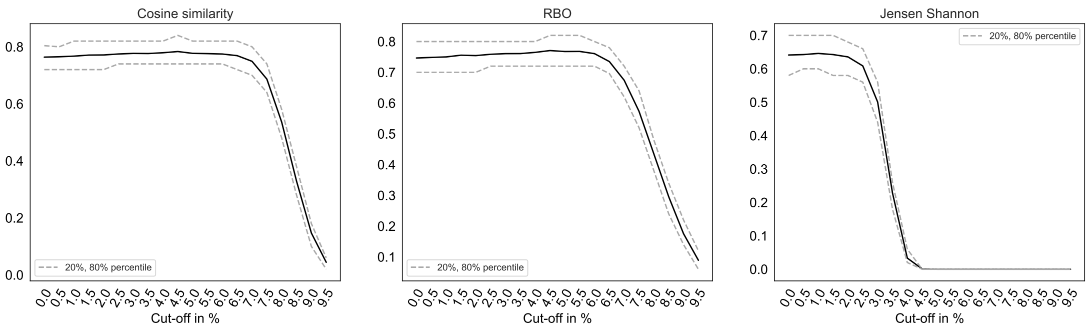
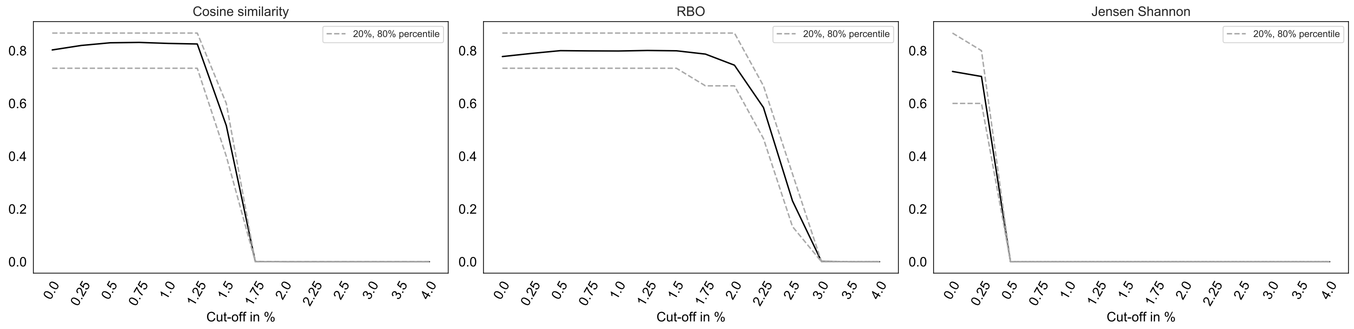


Figure A8: Recall values for DGP2 (higher values are preferred)


Cosine similarity
RBO
Jensen Shannon
20%, 80% percentile
Cut-off in %