Improving Latent Dirichlet Allocation: On Reliability of the Novel Method LDAPrototype

Jonas Rieger(✉), Jörg Rahnenführer, and Carsten Jentsch

Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
{rieger,rahnenfuehrer,jentsch}@statistik.tu-dortmund.de

Abstract. A large number of applications in text data analysis use the Latent Dirichlet Allocation (LDA) as one of the most popular methods in topic modeling. Although the instability of the LDA is mentioned sometimes, it is usually not considered systematically. Instead, an LDA is often selected from a small set of LDAs using heuristic means or human codings. Then, conclusions are often drawn based on the to some extent arbitrarily selected model. We present the novel method LDAPrototype, which takes the instability of the LDA into account, and show that by systematically selecting an LDA it improves the reliability of the conclusions drawn from the result and thus provides better reproducibility. The improvement coming from this selection criterion is unveiled by applying the proposed methods to an example corpus consisting of texts published in a German quality newspaper over one month.

Keywords: Topic model · Machine learning · Similarity · Stability · Stochastic

1 Introduction

Due to the growing number and especially the increasing amount of unstructured data, it is of great interest to be able to analyze them. Text data is an example for unstructured data and at the same time it covers a large part of them. It is organized in so-called corpora, which are given by collections of texts.

For the analysis of such text data topic models in general and the Latent Dirichlet Allocation in particular is often used. This method has the weakness that it is unstable, i.e. it gives different results for repeated runs. There are various approaches to reduce this instability. In the following, we present a new method LDAPrototype that improves the reliability of the results by choosing a center LDA. We will demonstrate this improvement of the LDA applying the method to a corpus consisting of all articles published in the German quality newspaper Süddeutsche Zeitung in April 2019.

E. Métais et al. (Eds.): NLDB 2020, LNCS 12089, pp. 118–125, 2020.
https://doi.org/10.1007/978-3-030-51310-8_11

1.1 Related Work

The Latent Dirichlet Allocation [3] is very popular in text data analysis. Numerous extensions to Latent Dirichlet Allocation have been proposed, each customized for certain applications, as the Author-Topic Model [18], Correlated Topics Model [2] or the more generalized Structural Topic Model [17]. We focus on LDA as one of the most commonly used topic models and propose a methodology to increase reliability of findings drawn from the results of LDA.

Reassigning words to topics in the LDA is based on conditional distributions, thus it is stochastic. This is rarely discussed in applications [1]. However, several approaches exist to encounter this problem based on a certain selection criterion. One of these selection criteria is perplexity [3], a performance measure for probabilistic models to estimate how well new data fit into the model [18]. As an extension, Nguyen et al. [13] proposed to average different iterations of the Gibbs sampling procedure to achieve an increase of perplexity. In general, it was shown that optimizing likelihood-based measures like perplexity does not select the model that fits the data best regarding human judgements. In fact, these measures are negatively correlated with human judgements on topic quality [5]. A better approach should be to optimize semantic coherence of topics as Chang et al. [5] proposed. They provide a validation technique called Word or Topic Intrusion which depends on a coding process by humans. Measures without human interaction, but almost automated, and also aiming to optimize semantic coherence can be transferred from the Topic Coherence [12]. Unfortunately, there is no validated procedure to get a selection criterion for LDA models from this topic's "quality" measure. Instead, another option to overcome the weakness of instability of LDA is to start the first iteration of the Gibbs sampler with reasonably initialized topic assignments [11] of every token in all texts. One possibility is to use co-occurences of words. The initialization technique comes with the drawback of restricting the model to a subset of possible results.

1.2 Contribution

In this paper, we propose an improvement of the Latent Dirichlet Allocation through a selection criterion of multiple LDA runs. The improvement is made by increasing the reliability of results taken from LDA. This particular increase is obtained by selecting the model that represents the center of the set of LDAs best. The method is called LDAPrototype [16] and is explained in Sect. 3. We show that it generates reliable results in the sense that repetitions lie in a rather small sphere around the overall centered LDA, when applying the proposed methods to an example corpus of articles from the Süddeutsche Zeitung.

2 Latent Dirichlet Allocation

The method we propose is based on the LDA [3] estimated by a Collapsed Gibbs sampler [6], which is a probabilistic topic model that is widely used in text data

analysis. The LDA assumes that there is a topic distribution for every text, and it models them by assigning one topic from the set of topics $T = \{T_1, ..., T_K\}$ to every token in a text, where $K \in \mathbb{N}$ denotes the user-defined number of modeled topics. We denote a text (or document) of a corpus consisting of M texts by

$$\boldsymbol{D}^{(m)} = \left(W_1^{(m)}, ..., W_{N^{(m)}}^{(m)}\right), \ m = 1, ..., M, \quad W_n^{(m)} \in \boldsymbol{W}, \ n = 1, ..., N^{(m)}.$$

We refer to the size of text m as $N^{(m)}$; $\boldsymbol{W} = \{W_1, ..., W_V\}$ is the set of words and $V \in \mathbb{N}$ denotes the vocabulary size. Then, analogously the topic assignments of every text m are given by

$$\boldsymbol{T}^{(m)} = \left(T_1^{(m)}, ..., T_{N^{(m)}}^{(m)}\right), \ m = 1, ..., M, \qquad T_n^{(m)} \in T, \ n = 1, ..., N^{(m)}.$$

Each topic assignment $T_n^{(m)}$ corresponds to the token $W_n^{(m)}$ in text m. When $n_k^{(mv)}, k = 1, ..., K, v = 1, ..., V$ describes the number of assignments of word v in text m to topic k, we can define the cumulative count of word v in topic k over all documents by $n_k^{(\bullet v)}$. Then, let $\boldsymbol{w}_k = (n_k^{(\bullet 1)}, ..., n_k^{(\bullet V)})^T$ denote the vectors of word counts for the $k = 1, ..., K$ topics. Using these definitions, the underlying probability model of LDA [6] can be written as

$$\begin{aligned} W_n^{(m)} \mid T_n^{(m)}, \boldsymbol{\phi}_k &\sim \text{Discrete}(\boldsymbol{\phi}_k), & \boldsymbol{\phi}_k &\sim \text{Dirichlet}(\eta), \\ T_n^{(m)} \mid \boldsymbol{\theta}_m &\sim \text{Discrete}(\boldsymbol{\theta}_m), & \boldsymbol{\theta}_m &\sim \text{Dirichlet}(\alpha), \end{aligned}$$

where α and η are Dirichlet distribution hyperparameters and must be set by the user. Although the LDA permits α and η to be vector valued [3], they are usually chosen symmetric because typically the user has no a-priori information about the topic distributions $\boldsymbol{\theta}$ and word distributions $\boldsymbol{\phi}$. Increasing η leads to a loss of homogenity of the mixture of words per topic. In contrast, a decrease leads to a raise of homogenity, identified by less but more dominant words per topic. In the same manner α controls the mixture of topics in texts.

3 LDAPrototype

The Gibbs sampler in the modeling procedure of the LDA is sensitive to the random initialization of topic assignments as mentioned in Sect. 1.1. We present a method that reduces the stochastic component of the LDA. This adaption of the LDA named LDAPrototype [16] increases the reliability of conclusions drawn from the resulting prototype model, which is obtained by selecting the model that seems to be the most central of (usually around) 100 independently modeled LDA runs. The procedure can be compared to the calculation of the median in the univariate case.

The method makes use of topic similarities measured by the modified Jaccard coefficient for the corresponding topics to the word count vectors $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$

$$J_m(\boldsymbol{w}_i, \boldsymbol{w}_j) = \frac{\sum_{v=1}^{V} \mathbb{1}_{\left\{n_i^{(\bullet v)} > c_i \ \wedge \ n_j^{(\bullet v)} > c_j\right\}}}{\sum_{v=1}^{V} \mathbb{1}_{\left\{n_i^{(\bullet v)} > c_i \ \vee \ n_j^{(\bullet v)} > c_j\right\}}},$$

where $\boldsymbol{c}$ is a vector of lower bounds. Words are assumed to be relevant for a topic if the count of the word passes this bound. The threshold $\boldsymbol{c}$ marks the modification to the traditional Jaccard coefficient [8] and can be chosen in an absolute or relative manner or as a combination of both.

The main part of LDAPrototype is to cluster two independent LDA replications using Complete Linkage [7] based on the underlying topic similarities of those two LDA runs. Let G be a pruned cluster result composed by single groups g consisting of topics and let $g_{|1}$ and $g_{|2}$ denote groups of g restricted to topics of the corresponding LDA run. Then, the method aims to create a pruning state where $g_{|1}$ and $g_{|2}$ are each build by only one topic for all $g \in G$. This is achieved by maximizing the measure for LDA similarity named S-CLOP (**S**imilarity of Multiple Sets by **C**lustering with **Lo**cal **P**runing) [16]:

$$\text{S-CLOP}(G) = 1 - \frac{1}{2K} \sum_{g \in G} |g| \left(||g_{|1}| - 1| + ||g_{|2}| - 1| \right) \in [0, 1].$$

We denote the best pruning state by $G^* = \arg\max\{\text{S-CLOP}(G)\}$ for all possible states G and determine similarity of two LDA runs by S-CLOP(G^*). The prototype model of a set of LDAs then is selected by maximizing the mean pairwise similarity of one model to all other models.

The methods are implemented in the R [14] package ldaPrototype [15]. The user can specify the number of models, various options for $\boldsymbol{c}$ including a minimal number of relevant words per topic as well as the necessary hyperparameters for the basic LDA α, η, K and the number of iterations the Gibbs sampler should run. The package is linked to the packages lda [4] and tosca [10].

4 Analysis

We show that the novel method LDAPrototype improves the Latent Dirichlet Allocation in the sense of reliability. To prove that, the following study design is applied to an example corpus from the German quality newspaper Süddeutsche Zeitung (SZ). The corpus consists of all 3 718 articles published in the SZ in April 2019. It is preprocessed using common steps for cleaning text data including duplicate removal leading to 3 468 articles. Moreover, punctuation, numbers and German stopwords are removed. In addition, all words that occur ten times or less are deleted. This results in $M = 3\,461$ non-empty texts and a vocabulary size of $V = 11\,484$. The preprocessing was done using the R package tosca [10].

4.1 Study Design

The study is as follows: First of all, a large number N of LDAs is fitted. This set represents the basic population of all possible LDAs in the study. Then we repeat P times the random selection of R LDAs and calculate their LDAPrototype. This means, finally P prototypes are selected, each based on R basic LDAs, where each LDA is randomly drawn from a set of N LDAs. Then, a single prototype is

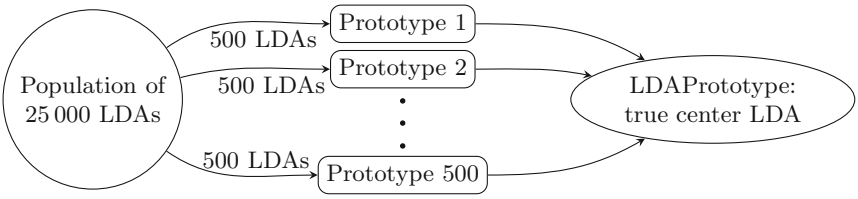


Fig. 1. Schematic representation of the study design for $N = 25\,000$ LDAs in the base population and $P = 500$ selected prototypes, each based on $R = 500$ sampled LDAs from the base population.

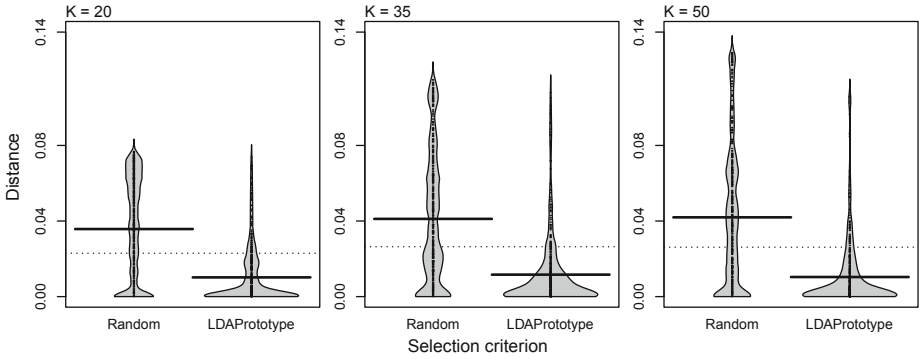
determined based on a comparison of the P prototypes. This particular prototype forms the assumed true center LDA. In addition, we establish a ranking of all other prototypes. The order is determined by sequentially selecting the next best prototype which realizes the maximum of the mean S-CLOP values by adding the corresponding prototype and simultaneously considering all higher ranked LDAPrototypes.

For the application we choose three different parameter combinations for the basic LDA. In fact, we want to model the corpus of the SZ with $K = 20, 35, 50$ topics. We choose accordingly $\alpha = \eta = 1/K$ and let the Gibbs sampler iterate 200 times. We choose the size of the population as $N = 25\,000$, so that we initially calculate a total of 75 000 LDAs, which is computationally intensive but bearable. We use the R package ldaPrototype [15] to compute the models on batch systems. We set the parameters of the study to a sufficiently high and at the same time calculable value of $P = R = 500$. That is, we get 500 PrototypeLDAs, each based on 500 basic LDAs, that are sampled without replacement from the set of 25 000 basic LDAs. The sampling procedure is carried out without replacement in order to protect against falsification by multiple selection of one specific LDA. Figure 1 represents this particular study design schematically.

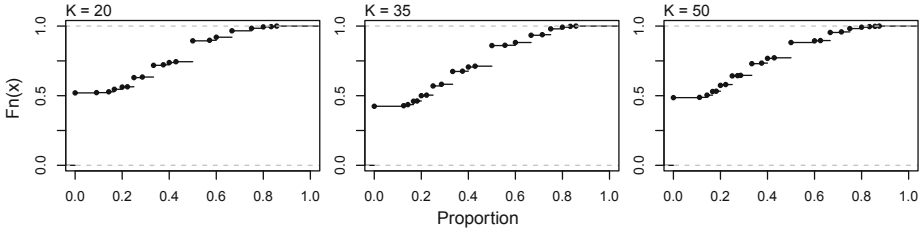
Then, we inspect the selection of the P prototypes. On the one hand, we quantify the goodness of selection by determining how many LDAs, that were available in the corresponding run, are ranked before the corresponding LDAPrototype. On the other hand, the analysis of the distance to the best available LDA run in the given prototype run provides a better assessment of the reliability of the method. We compare the observed values with randomized choices of the prototype. This leads to statements of the form that the presented method LDAPrototype selects its prototypes only from a sufficiently small environment around the true center LDA, especially in comparison to random selected LDAs.

4.2 Results

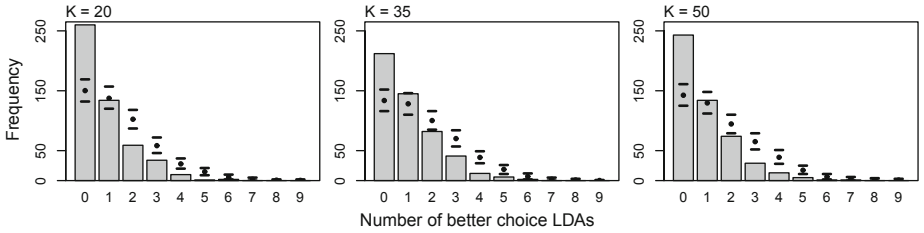
For the analysis we first determine the true center LDA and a ranking for all 500 prototypes as described in Sect. 4.1 for each $K = 20, 35, 50$. The corresponding mean S-CLOP value at the time of addition is assigned to each prototype in the ranking as a measure of proximity to the true center LDA. To visualize the



(a) Distance of each of the LDAPrototypes to the LDA that would have been the best choice in the corresponding prototype run regarding closeness to the center LDA.



(b) Empirical cumulative distribution function of the proportion of how many LDAs are closer to the center LDA than the selected LDAPrototype.



(c) Number of LDAs that are closer to the center LDA than the selected LDAPrototype.

Fig. 2. Analysis of the improvement of reliability by using the LDAPrototype for $K = 20, 35, 50$ modeled topics. Every single value corresponds to one of the $P = 500$ prototype runs resulting in the corresponding LDAPrototype.

rankings, we use so-called beanplots [9] as a more accurate form of boxplots, as well as empirical cumulative distribution functions (ECDF) and bar charts.

For $K = 20, 35, 50$ each of the 25 000 LDAs is included at least once in the 500 times 500 selected LDAs. Nevertheless, only 169, 187 and 186 different LDAs are chosen as prototypes. The LDAPrototype method thus differs significantly from a random selection, whose associated simulated 95% confidence interval suggests between 490 and 499 different prototypes.

Figure 2 summarizes the analysis of the increase of reliability for 20, 35 and 50 topics, respectively. The beanplots in Fig. 2a indicate the distance of each LDA actually selected from the LDAPrototype method to the supposedly most suitable LDA from the identical prototype run with respect to the values from the ranking. For comparison, the distribution of the distances for random selection of the prototype is given besides. The corresponding values were generated by simulation with permutation of the ranking. The ECDFs in Fig. 2b show the relative number of LDAs, in each of the $P = 500$ prototype runs, that according to the ranking would represent a better choice as prototype. Finally, the bar charts in Fig. 2c show the corresponding distribution of the absolute numbers of available better LDAs in the same run in accordance to the determined ranking of prototypes. In addition, simulated 95% confidence intervals for frequencies realized by the use of random selection are also shown.

For $K = 20$, many randomly selected LDAs have a rather large distance of about 0.07 at a total mean value of just below 0.04, while the presented method realizes distances that are on average below 0.01. For increasing K the distances seem to increase as well. While the random selection produces an almost unchanging distribution over an extended range, the distribution of LDAPrototype shifts towards zero. Higher values become less frequent. The ECDFs look very similar for all K, whereby for $K = 35$ slightly lower values are observed for small proportions. This is supported by the only major difference in the bar charts. Modeling 20 or 50 topics, for 50% of the prototype runs there is no better available LDA to choose, while for the modeling of 35 topics this scenario applies for just over 40%. The corresponding confidence intervals in Fig. 2c are lowered as well. This is an indication that for $K = 35$ it is easier to find a result that is stable to a certain extent for the basic LDA. This is supported by the fact that the distribution of distances in Fig. 2a does not seem to suffer.

5 Discussion

We show that the LDAPrototype method significantly improves the reliability of LDA results compared to a random selection. The presented method has several advantages, e.g. the automated computability, as no need of manual coding procedures. In addition, besides the intuitive statistical approach, the proposed method preserves all components of an LDA model, especially the specific topic assignments of each token in the texts. This means that all analyses previously carried out on individual runs can be applied to the LDAPrototype as well. The results suggest that $K = 35$ topics produces more stable results and might therefore be a more appropriate choice for the number of topics than $K = 20$ or 50 on the given corpus. Further studies to analyze the observed differences in the number of better LDAs as well as the distances to the best LDA between different choices of the numbers of topics, may lead to progress in the field of hyperparameter tuning for the LDA.

References

1. Agrawal, A., Fu, W., Menzies, T.: What is wrong with topic modeling? And how to fix it using search-based software engineering. Inf. Softw. Technol. **98**, 74–88 (2018). https://doi.org/10.1016/j.infsof.2018.02.005
2. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. Ann. Appl. Stat. **1**(1), 17–35 (2007). https://doi.org/10.1214/07-AOAS114
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/10.1162/jmlr.2003.3.4-5.993
4. Chang, J.: LDA: Collapsed Gibbs Sampling Methods for Topic Models (2015). https://CRAN.R-project.org/package=lda. R package version 1.4.2
5. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Proceedings of the 22nd International NIPS-Conference, pp. 288–296. Curran Associates Inc. (2009)
6. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(Suppl. 1), 5228–5235 (2004). https://doi.org/10.1073/pnas.0307752101
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
8. Jaccard, P.: The distribution of the flora in the alpine zone. New Phytol. **11**(2), 37–50 (1912). https://doi.org/10.1111/j.1469-8137.1912.tb05611.x
9. Kampstra, P.: Beanplot: a boxplot alternative for visual comparison of distributions. J. Stat. Softw. Code Snippets **28**(1), 1–9 (2008). https://doi.org/10.18637/jss.v028.c01
10. Koppers, L., Rieger, J., Boczek, K., von Nordheim, G.: tosca: Tools for Statistical Content Analysis (2019). https://doi.org/10.5281/zenodo.3591068. R package version 0.1-5
11. Maier, D., et al.: Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun. Methods Measur. **12**(2–3), 93–118 (2018). https://doi.org/10.1080/19312458.2018.1430754
12. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 EMNLP-Conference, pp. 262–272. ACL (2011)
13. Nguyen, V.A., Boyd-Graber, J., Resnik, P.: Sometimes average is best: the importance of averaging for prediction using MCMC inference in topic modeling. In: Proceedings of the 2014 EMNLP-Conference, pp. 1752–1757. ACL (2014). https://doi.org/10.3115/v1/D14-1182
14. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). http://www.R-project.org/
15. Rieger, J.: LDAPrototype: Prototype of Multiple Latent Dirichlet Allocation Runs (2020). https://doi.org/10.5281/zenodo.3604359. R package version 0.1.1
16. Rieger, J., Koppers, L., Jentsch, C., Rahnenführer, J.: Improving Reliability of Latent Dirichlet Allocation by Assessing Its Stability Using Clustering Techniques on Replicated Runs (2020)
17. Roberts, M.E., Stewart, B.M., Tingley, D., Airoldi, E.M.: The structural topic model and applied social science. In: NIPS-Workshop on Topic Models: Computation, Application, and Evaluation (2013)
18. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th UAI-Conference, pp. 487–494. AUAI Press (2004)