Evaluating Hyperparameter Alpha of LDA Topic Modeling

Du, Keli
duk@uni-trier.de
Universität Trier, Germany

Introduction

As a quantitative text analysis method, Latent Dirichlet Allocation (LDA), also often referred to as topic modeling (Blei 2012), has been widely used in Digital Humanities in recent years to explore numerous unstructured text data. When topic modeling is used, one has to deal with many parameters that could influence the result, such as the hyperparameter Alpha and Beta, the topic number, document length, or the number of iterations when updating the model. To understand the impact of these parameters, they must be systematically evaluated. In the last few years, there have been several studies evaluating LDA topic modeling in Digital Humanities or Computational Literary Studies (e.g., Jockers 2013; Schöch 2017; Du 2020; Uglanova & Gius 2020) and the presented paper focuses on evaluating the impact of hyperparameter Alpha on LDA topic models.

Hyperparameter Alpha can refer to two different types of parameters in the context of LDA topic modeling: LDA model parameter and inference algorithm parameter. As a parameter of the LDA model, Alpha determines the properties of a Dirichlet distribution, which is the prior probability distribution of the topic-document distribution. Together, the hyperparameter Alpha and the prior probability distribution determine which topics we expect to occur more frequently in the corpus and how confident we are about them. In practice, when we employ Gibbs Sampling to train our topic model, Alpha is the parameter, which has the smoothing effect on the topic-document distribution and ensures that the probability of each topic in each document is not 0 throughout the entire inference procedure. More importantly, Alpha represents the assumption about the data on how topics are distributed in documents before inferencing the topic model. In other words, the hyperparameter Alpha affects how often each topic occurs in each document. When the alpha value of a topic is set larger in a document, it means that the topic has a greater chance of appearing in that document. And vice versa. For this reason, the setting of Alpha can affect the quality of the inferenced topic model. Therefore, this paper focuses on evaluating the impact of inference algorithm parameter alpha systematically.

According to Griffiths & Steyvers (2004), the topic model has the best quality when the sum of Alphas of all topics is equal to 50. This is probably the reason that in MALLET 2.0.7, the default value of the sum of Alphas was set to 50, while in MALLET 2.0.8, the value is reduced to 5. According to the supervisor of MALLET, David Mimno: “The general experience was that 50 was too large, and that 5 is a better default.”[1] Since there are different opinions on this issue, it is interesting to test how Alpha affects LDA topic modeling, especially on different types of text collections that are not in English. Therefore, this paper presents a study on evaluating Alpha on two German text collections and aims to understand the influence of hyperparameter Alpha from two perspectives: topic modeling based single-label document classification and topic coherence, representing the quality of the topic model and the quality of the topics, respectively.

Method

Two collections of German texts were built for the study. The first corpus is a collection of 2000 newspaper articles published between 2001 and 2014. The articles belong to ten different thematic classes, and each class contains 200 articles. The ten classes are “Digital”, “Society”, “Career”, “Culture”, “Lifestyle”, “Politics”, “Travel”, “Sports”, “Study” and “Economy”. The corpus contains over 3.4 million words in total, and the average text length is about 1800 words. The second corpus consists of 439 dime novels published between 1961 and 2016, and they belong to five subgenres, namely 100 fantasy novels, 51 horror novels, 88 crime novels, 100 romance novels, and 100 science fiction stories. The corpus contains about 13.4 million words, and the average text length is about 30,000 words. All texts are lemmatized. Since the average document length of the newspaper articles is 1800 words, the novels are also split into 1800 words segments. Thus, the document length is no longer a confounding factor when comparing the test results on the newspaper corpus with the results on the novel corpus.

The goal of the following tests is to explore the influence of the hyperparameter Alpha. While training topic models, the setting is varied by the value of Alpha $\in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 20, 30, 40, 50, 100\}$ and number of topics $\in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500\}$. For all other parameter settings, the default values of the topic modeling software were taken. All models were trained without applying hyperparameter optimization, which means that if Alpha is set to 0.1, the Alpha value for each topic is set to 0.1 during the whole training process. Common stop words were removed from both corpora. For technical reasons, namely random initialization in the topic assignment and Gibbs sampling, two topic models from one corpus are not completely identical even if the parameter settings during training are the same. Therefore, ten models were trained for each setting to balance the randomness from the technical side.

The topic models were trained using MALLET (McCallum 2002). As a result, a topic-document distribution and the topics are obtained for each topic model. In a topic-document distribution, each document is represented by an N -dimensional vector, while N is the number of topics of the topic model. Based on the topic-document distribution, the document classification was performed, and the classification was done as a 10-fold cross-validation with a linear SVM classifier. For the newspaper corpus, the articles were classified according to their thematic classes. For the novel corpus, the novel segments were classified according to their subgenre. The topic coherence was automatically calculated by the Java program Palmetto (Röder et al. 2015), and the first ten most important words of each topic were taken for the calculation. The reference corpus for the calculation of the topic coherence is the lemmatized German Wikipedia. Several topic coherence mea-sures have been implemented in Palmetto. For this work, the Nor-malized Pointwise Mutual Information (NPMI) based coherence measure proposed in Aletras & Stevenson (2013) was taken. The theoretical range of NPMI based coherence measure is between -1 and 1. The higher the score, the better the topic.

By performing Bag-of-Words (BoW) model-based classification, a baseline of document classification has been defined for both corpora. The tests were also done as a 10-fold cross-valida-

tion with a linear SVM classifier. The F1(macro) score for the newspaper articles and for the novel segments was 0.758 and 0.993, respectively. A baseline of the NPMI value was also defined for each corpus. With only one iteration, 14 topic models were first trained on each corpus, containing 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 topics, respectively. In this way, 1,950 “topics before topic modeling” have been trained for each corpus. The NPMI scores of these topics were then calculated, and the average NPMI score is the NPMI baseline, which is -0.0619 for the newspaper corpus and -0.1153 for the novel corpus. A black line represents the baselines in Figure 3 and Figure 4.

Results

Document classification : Figure 1 and Figure 2 show the classification results based on topic models of newspaper articles and novel segments, respectively. It can be seen in both figures that the classification results gradually become worse with the increase of the setting of Alpha, regardless of how many topics have been trained. Especially if Alpha is set to greater than 1, the classification results based on topic models with more topics (the blue lines) show a stronger decreasing trend than the results based on topic models with fewer topics (the red lines). In comparison, most F1 scores change less when Alpha is set to a value smaller than 1. However, we can still see that the blue lines start to decrease when the Alpha is raised from 0.5 to 1. The highest F1-score of classifying newspaper articles and novel segments in this test are 0.752 and 0.998, respectively, which do not differ much from the predefined baseline based on BoW-model.

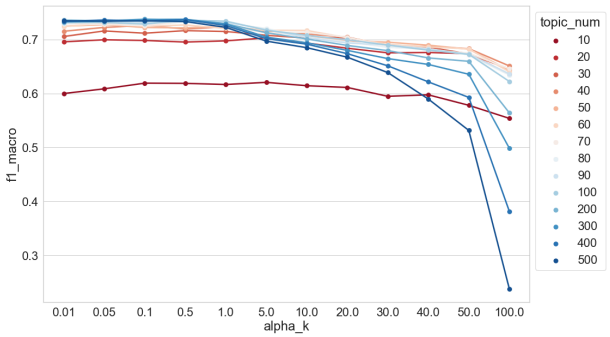

Fig. 1: Average F1(macro)-scores of topic modeling based classification of newspaper articles

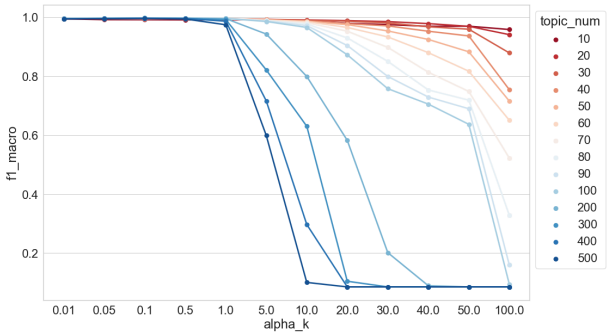

Fig. 2: Average F1(macro)-scores of topic modeling based classification of novel segments

Topic coherence: Compared to the classification results, the evaluation from the perspective of topic coherence shows some differences between the two corpora. Firstly, it can be observed in Figure 3 that the maximum of the NPMI-score distributions decreases with the increase of Alpha from almost 0.3 to about 0.12. In addition, the median of the distributions also shows a decreasing trend. At Alpha = 0.01, the median is lower than the NPMI baseline if the number of topics is set higher than 100. However, at Alpha = 100, the median is already lower than the NPMI baseline if the number of topics is set to 70. Apart from that, we can observe that the topic models with a higher number of topics contain more topics with low NPMI scores, regardless of the setting of Alpha. Compared to the test on the newspaper corpus, the test results on the novel corpus are slightly different. When Alpha is set smaller than 1, the NPMI-score distributions do not show an evident change as Alpha increases, and the range of distribution is often broader when the number of topics is set between 60 and 300. Starting from Alpha being raised to greater than 1, the distributions of the NPMI-scores clearly change, and the results then are similar to the previous test on the newspaper corpus: the maximum of the NPMI-score distributions decreases with the increase of the Alpha, and topic models with a higher number of topics contain more topics with low NPMI scores.

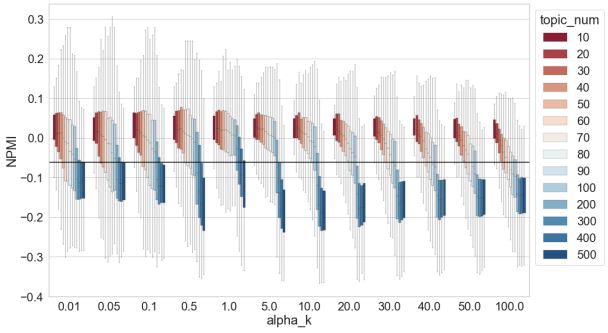

Fig. 3: NPMI-score distributions of topics from newspaper articles

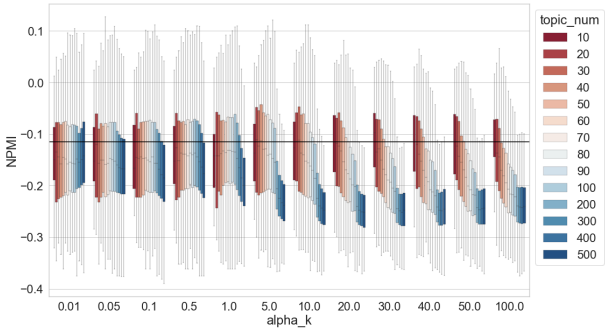

Fig. 4: NPMI-score distributions of topics from novel segments

Conclusion

The presented research evaluates the influence of hyperparameter Alpha in topic modeling on a German newspaper corpus and a German literary text corpus from two perspectives, single-label document classification, and topic coherence. Based on the results of the presented investigation, it can be stated that one should

avoid training topic models with a setting of the Alpha of each topic to greater than 1 in order to ensure better topic modeling based document classification results and to get more coherent topics. In addition to that, LDA topic models with many topics are more vulnerable to changes in Alpha. Therefore, with the result of the presented investigations in this study, one can confirm the explanation of Mimno mentioned earlier that a smaller Alpha is better suitable for LDA Topic Modeling.

Footnotes

1. https://stackoverflow.com/questions/45162186/mallet-topic-modeling-topic-keys-output-parameter (15.07.2021)

Bibliography

Aletras, Nikolaos / Stevenson, Mark (2013): “Evaluating topic coherence using distributional semantics”. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (pp. 13-22).

Blei, David M. (2012): “Probabilistic topic models”, in: *Communications of the ACM* , 55(4), 77-84.

Du, Keli (2020): „Der Spielraum zwischen "zu wenig" und "zu viel"“. Presented at the DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. 7. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2020), Paderborn: Zenodo. http://doi.org/10.5281/zenodo.4621770 .

Griffiths, Thomas L. / Steyvers, Mark (2004): “Finding scientific topics”, in: *Proceedings of the National Academy of Sciences, 101* (Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101 .

Jockers, Matthew L. (2013): *Macroanalysis: Digital methods and literary history* . University of Illinois Press.

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu .

Röder, Michael / Both, Andreas / Hinneburg, Alexander (2015): “Exploring the space of topic coherence measures”, in: *Proceedings of the eighth ACM international conference on Web search and data mining* , 399–408.

Schöch, Christof (2017): “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama.”, in: *Digital Humanities Quarterly* 11, no. 2. §1-53. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html .

Uglanova, Inna / Gius, Evelyn (2020): “The Order of Things. A Study on Topic Modelling of Literary Texts”, in: *Online Workshop on Computational Humanities Research, Proceedings* . http://ceur-ws.org/Vol-2723/long7.pdf .