Research About Latent Dirichlet Allocation

Zhibo Liang

zbliang_sasee@buaa.edu.cn

Abstract. Topic modeling is one of the most powerful techniques in text mining for data mining, latent data discovery, and finding relationships among data and text documents. This report performs a research about Latent Dirichlet Allocation model on Chinese Text. A book named "Bixuejian" is ultilized to operate the calculating and plotting. As a result, the performance of model is acquired and perplexity is applied to evaluate the performance.

Keywords: Topic modeling, Latent Dirichlet allocation

1 Introduction

Topic modeling methods are powerful smart techniques that widely applied in natural language processing to topic discovery and semantic mining from unordered documents[1]. In a wide perspective, Topic modeling methods based on Latent Dirichlet Allocation(LDA) have been applied to natural language processing, text mining, and social media analysis, information retrieval. LDA are applied in various fields including medical sciences, software engineering, geography, political science, etc[2]. Thus, a research on LDA is performed in this report.

2 Methodology

2.1 Latent Dirichlet Allocation

The basic idea of LDA is to assume that each document is composed of a set of topics, and each topic is composed of a set of words. By analyzing the frequency of words in the documents, LDA can infer the underlying distribution of topics. The core proceeding of LDA can be detailed as follows:

$$p(\omega_i|\alpha,\beta) = \int_{\theta_i}\int_{\Phi}\sum_{z_i} p(\omega_i, z_i, \theta_i, \Phi|\alpha,\beta) \tag{1}$$

where ω_i represents the generated word or char; α and β is Dirichlet Distribution.

2.2 Perplexity

Perplexity is a metric used to evaluate language models or probabilistic models, measuring the model's ability to predict new samples. A lower perplexity indicates that the model can better predict new data and has a better understanding of the underlying patterns. Perplexity is commonly used in natural language processing tasks such as machine translation, speech recognition, and text generation. It is calculated based on the probability distribution assigned by the model to the predicted outcomes. The perplexity formula in sequence models can be derived as follows:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1, w_2, \ldots, w_{i-1})\right) \tag{2}$$

And in topic models, it can be explained as:

$$\text{Perplexity} = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\log p(w_{d,n})}{\sum_{d=1}^{D} N_d} \tag{3}$$

where D represents the total number of documents in the corpus; N_d represents the number of words in document d; $w_{d,n}$ represents the nth word in document d; $p(w_{d,n})$ is the probability of the word $w_{d,n}$ according to the topic model.

3 Experiment Results

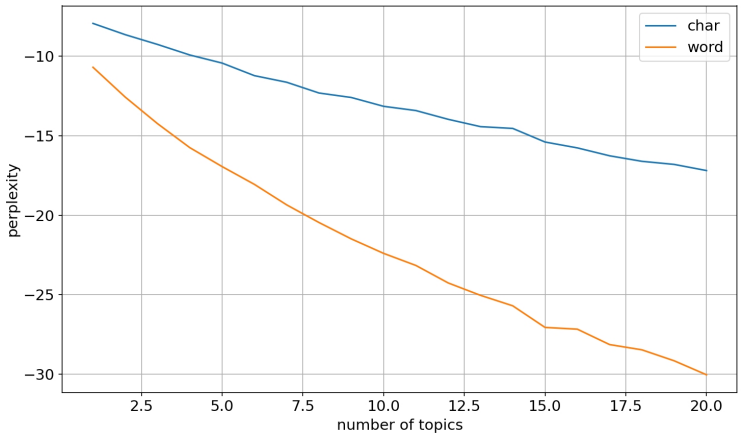

Fig. 1: The perplexity-topics curve.

To ensure credibility, the results of this experiment were obtained using ten-fold cross-validation. The experiment is based on book "Bixuejian" and 1000

tokens are applied in each situation. Besides, the number of iterations for each experiment is set to 20.

3.1 The impact of different number of topics

The number of topics is an important parameter in topic modeling algorithms, especially Latent Dirichlet Allocation (LDA). It determines the granularity and specificity of the discovered topics. The choice of the number of topics in LDA is a trade-off between capturing meaningful themes and avoiding overfitting or underfitting. Through Fig.1, it can be observed that as the number of topics increases, the perplexity gradually decreases(choose K=20).

3.2 The impact of different token lengths

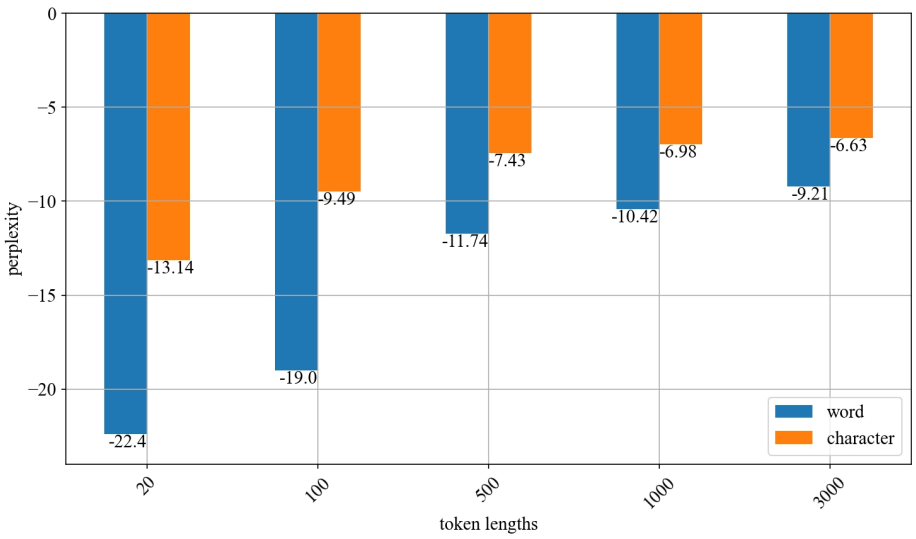

Fig. 2: The impact of different token lengths on perplexity.

The length of tokens can also have an influence on various aspects of the analysis and modeling process. Longer texts may contain more topic information, which can help the model learn more accurate topic distributions. On the other hand, shorter texts may have less topic information, leading to increased difficulty for the model to learn effectively. And in Fig.2, the perplexity increases(choose T=10) as token lengths increases.

4 Conclusion

The results of experiments indicate that the impact of topic number and token lengths is obvious. The perplexity of word-level LDA is lower than that of char level model. Therefore, in the future research, it is crucial for us to select appropriate parameters.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
2. D. Jiang, J. Vosecky, K. W.-T. Leung, L. Yang, and W. Ng, "Sg-wstd: A framework for scalable geographic web search topic discovery," *Knowledge-Based Systems*, vol. 84, pp. 18–33, 2015.