Report of LDA Model

Zhenyang Zhu
zhenyangzhu@buaa.edu.cn

Abstract

This is the third assignment report for the NLP course. This report conducted LDA topic modeling and classification experiments on the Jin Yong's novel corpus, exploring the impact of the number of topics K on model construction and classification. It also compared the classification results based on character-level and word-level units. The results showed that as the number of topics K increased, the classification accuracy generally increased, but overfitting occurred when K was too large. Meanwhile, the classification results based on character-level units were more accurate than those based on word-level units.

Methodology

Latent Dirichlet Allocation (LDA) is a widely used probabilistic topic modeling technique, which aims to discover the hidden themes or topics that are present in a collection of documents. LDA is an unsupervised learning method, meaning that it does not require labeled data to perform the analysis. Instead, LDA extracts latent topics from the corpus of text, and each document is represented as a mixture of these topics.

In the field of natural language processing, the LDA model is a generative probabilistic model that represents each document as a distribution over a fixed set of topics, where each topic is a distribution over a fixed set of words. The LDA model can be viewed as a method for clustering documents into groups based on the presence of certain topics.

The LDA model assumes that each document in a corpus is a mixture of various topics, and that each topic is characterized by a distribution over words. Specifically, the LDA model represents a document as a probability distribution over topics, and each topic as a probability distribution over words.

Mathematically, the LDA model can be formulated as follows:

Given a corpus of **D** documents, each document **d** is represented as a bag-of-words, **i.e.**, a vector of word counts $w_d = (w_{d,1}, w_{d,2}, ..., w_{d,V})$, where V is the size of the vocabulary. Let K be the number of topics in the corpus, and let α and β be hyperparameters of the model.

The LDA model defines the joint probability distribution of the corpus as follows:

$$P(w, z, \theta, \phi \mid \alpha, \beta) = \prod_{d=1}^{D} P(\theta_d \mid \alpha) \prod_{n=1}^{N_d} P(z_{d,n} \mid \theta_d) P(w_{d,n} \mid z_{d,n}, \phi_{z_{d,n}})$$

where $z_{d,n}$ is the topic assigned to the nth word in document **d**, θ_d is the topic distribution of document **d**, ϕ_k is the word distribution of topic k, and N_d is the number of words in

document **d**.

The LDA model assumes the following generative process:

1. For each topic $k = 1, ..., K$, draw a word distribution ϕ_k from a Dirichlet distribution with parameter β.

2. For each document $\boldsymbol{d} = 1, ..., D$, draw a topic distribution θ_d from a Dirichlet distribution with parameter α.

3. For each word $w_{d,n}$ in document **d**:

   a. Draw a topic $z_{d,n}$ from the document's topic distribution θ_d.

   b. Draw a word $w_{d,n}$ from the topic's word distribution $\phi_{z_{d,n}}$.

In other words, LDA assumes that each document is a mixture of K topics, where K is a pre-specified parameter. Each topic is represented by a distribution over the vocabulary of the corpus. The model assumes that the topic distribution for each document is drawn from a Dirichlet prior with parameter α, and the word distribution for each topic is drawn from a Dirichlet prior with parameter β.

The LDA model can be learned using Bayesian inference, specifically the Gibbs sampling algorithm. The goal of inference is to compute the posterior distribution over the latent variables z and ϕ given the observed corpus.

The goal of LDA is to infer the topic distribution for each document and the word distribution for each topic given the observed words in the corpus. This is achieved using Bayesian inference techniques, such as variational inference or Markov Chain Monte Carlo (MCMC) methods.

LDA has been successfully applied to a wide range of applications such as text classification, sentiment analysis, and recommendation systems. In text classification, for example, LDA can be used to automatically categorize documents into different topics. LDA can also be used in recommendation systems to suggest items to users based on their interests and preferences.

In conclusion, LDA is a powerful probabilistic model for topic modeling and has been successfully applied to a wide range of applications. The model assumes that each document is a mixture of a small number of topics and that each topic is a distribution over words. The goal of LDA is to infer the topic distribution for each document and the word distribution for each topic given the observed words in the corpus. This is achieved using Bayesian inference techniques, such as variational inference or MCMC methods.

Experimental Studies

This report uses the LDA model for text modeling on a corpus of 16 Jin Yong novels. 200 paragraphs are uniformly sampled from the given corpus, with each paragraph consisting of 500 words and labeled according to the novel it belongs to. The LDA model is used for text modeling, and each paragraph is represented as a distribution of topics for classification. The classification method used is the support vector machine (SVM) method.

The Python library gensim is used for LDA text modeling. Gensim is a simple and efficient Python library for natural language processing, used to extract semantic topics from documents. Gensim takes raw, unstructured text (plain text) as input, and its built-in algorithms include Word2Vec, FastText, and Latent Dirichlet Allocation (LDA). By computing statistical co-occurrence patterns in the training corpus, it automatically discovers the semantic structure of

the documents.

In the data preprocessing stage, it was found that the paragraph division in the provided corpus of Jin Yong novels is complex, and dividing it directly based on Chinese paragraphs is obviously impractical. Therefore, this report considers uniformly sampling 13 paragraphs from each of the 16 novels, for a total of 208 paragraphs, with each paragraph consisting of 500 words, to construct the training dataset. Additionally, for the convenience of testing in the future, the 501st to 1000th words were taken as the test set. Furthermore, in the process of reading the documents, Chinese stop words, punctuation marks, and meaningless contents such as "新语丝电子文库" that are inherent in the documents were also considered and deleted.

The LDA model contains hyperparameters such as the number of topics K, the topic distribution parameter α, and the word distribution parameter β for each topic. This report mainly considers the influence of the number of topics K on the model construction. It is worth noting that the corpus consists of 16 Jin Yong novels, so this report first chooses K=16 as the number of topics. After building the model, the top 6 most important words in each topic are outputted, as shown in **Table 1**.

Table 1 The words of each topic in LDA model with K=16

Topic						
1	道	范蠡	勾践	便	师兄	薛烛
2	道	说	便	中	听	一个
3	道	麽	说	李文秀	中	便
4	道	说	牋	中	便	麽
5	道	说	王夫人	黄眉僧	段誉	请
6	道	范蠡	便	说	阿青	爷爷
7	道	韦小宝	便	康熙	说	皇上
8	道	中	韦小宝	说	萧中慧	便
9	道	韦小宝	说	袁承志	康熙	便
10	道	说	便	中	见	听
11	韦小宝	公主	道	图尔布青	门闩	见
12	道	便	剑士	中	说	说道
13	道	大汉	袁承志	说	令狐冲	丁典
14	道	说	便	中	杨过	武功
15	道	说	中	便	袁承志	一个
16	道	万圭	说	中	便	见

Moreover, based on the topic distribution of each paragraph obtained from the LDA model, this report uses the support vector machine (SVM) method for classification. The labeled results obtained on the training and test sets are shown in **Figure 1**. Among them, the colored part is the correctly classified paragraph. The accuracy of classification on the training set is **25.96%**, and the accuracy on the test set is **25.48%**.

实际分类结果	SVM分类结果												
0	10	15	9	11	7	8	4	7	8	1	10	1	8
1	11	1	1	15	10	15	4	7	8	9	14	1	9
2	1	15	11	10	10	15	9	7	7	3	10	1	7
3	3	9	9	1	14	8	15	3	15	11	1	3	3
4	10	9	4	4	9	7	1	8	11	9	10	11	3
5	9	10	8	9	4	14	14	14	9	14	8	9	9
6	1	11	6	8	8	8	8	14	14	14	8	6	8
7	15	11	7	9	14	7	14	15	7	9	14	7	11
8	8	8	8	11	8	10	7	8	8	8	8	3	3
9	14	14	7	14	9	3	9	9	8	10	11	9	9
10	9	9	9	10	10	10	10	10	10	10	10	10	10
11	14	10	11	8	11	14	1	11	15	1	8	9	11
12	10	8	14	7	8	10	1	1	1	8	8	8	9
13	7	3	8	9	8	1	9	15	14	14	14	9	1
14	15	14	14	14	7	6	9	9	1	14	14	4	9
15	9	15	4	15	9	15	15	15	7	15	11	7	15

Fig 1 The classification result of training dataset with K=16

Subsequently, this report explores the impact of different numbers of topics K on the construction of the LDA model and the final classification. LDA models are constructed with K=1, 5, 10, 20, 50, 100, and 200, 500, respectively, and the words contained in each topic and the classification results are outputted. The result of K=50 is shown in **Table 2** and **Figure 2**.

Table 2 The words of top 20 topics in LDA model with K=50

Topic						
1	道	虚竹	洪七公	女童	松球	小龙女
2	道	说	便	中	韦小宝	胡斐
3	道	万震山	丁典	韦小宝	万圭	苏鲁克
4	高季兴	高氏	保勖	王超	荆南	念珠
5	道	麽	说	韦小宝	公主	少女
6	道	听	便	说	中	剑士
7	道	说	便	中	见	听
8	道	李文秀	麽	说	著	中
9	道	师父	说	黄眉僧	李莫愁	小龙女
10	道	曹云奇	二人	穆念慈	麽	著
11	道	李文秀	男孩	麽	天铃鸟	便
12	杨行密	张训	部下	杀	百姓	孙儒
13	道	勾践	薛烛	师兄	便	伍子胥
14	吴士	少女	吴	国剑士	八名	山羊
15	道	便	派	说	唐山人	中
16	道	说	那文士	小孩	说道	天下
17	中	袁崇焕	便	大师	道	二人
18	道	说	便	说道	令狐冲	中
19	大汉	袁承志	令狐冲	虚竹	小慧	见
20	韦小宝	吴三桂	九难	渤泥国	渤泥	白衣

实际分类结果	SVM分类结果												
0	10	14	5	0	0	6	0	13	0	4	0	3	8
1	4	0	3	1	5	12	1	7	6	10	10	1	14
2	0	2	2	2	3	0	5	14	4	11	12	4	2
3	3	3	3	10	14	3	3	3	15	4	3	3	6
4	12	5	3	0	5	4	4	4	4	4	10	12	4
5	5	5	5	5	0	9	5	13	5	3	12	1	5
6	6	3	6	6	6	6	6	12	11	5	6	6	10
7	15	4	7	7	7	7	7	11	7	5	1	2	3
8	8	3	3	2	4	13	14	8	6	7	8	3	8
9	2	9	12	14	9	3	9	13	0	10	6	10	6
10	15	5	7	1	10	10	10	10	10	10	10	10	10
11	14	43	11	8	8	5	7	11	11	3	11	14	11
12	7	12	12	9	6	10	12	12	1	12	12	12	12
13	0	7	9	13	6	12	13	3	9	12	12	4	13
14	14	9	9	14	13	12	12	11	14	8	9	14	4
15	15	13	9	15	1	14	15	14	15	1	11	7	3

Fig 2 The classification result of training dataset with K=50

The line chart of classification accuracy of the training and test sets under different values of K is shown in **Figure 3**. It can be seen that the accuracy generally increases with the increase of K. This is because as the number of topics increases, the model can better fit the characteristics of each paragraph, and the basis for classification also increases, resulting in more accurate classification. However, when K increases over 20, the classification accuracy decreases, indicating overfitting. To address this issue, one can reduce the number of iterations in the model appropriately, to decrease the occurrence of overfitting.

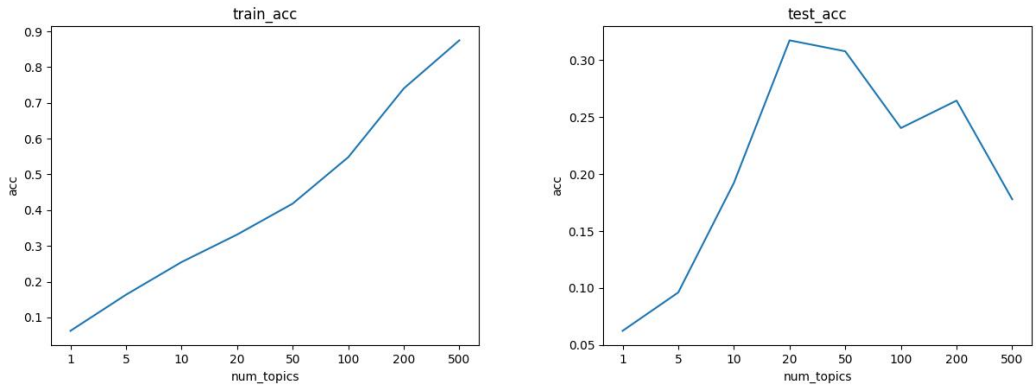

Fig 3 The curve of the accuracy of the training set and the test set as K increases

In addition, this report also explored the effect of using different basic units (characters or words) on the classification results in LDA topic modeling. The LDA model with 16 topics is used, and the topic distributions and classification results are presented in **Table 3** and **Figure 4,** respectively, for the case when characters are used as the basic unit. The accuracy of classification on the training set is **52.4%**, and the accuracy on the test set is **52.9%**. The results suggest that using characters as the basic unit results in a higher classification accuracy compared to using words as the basic unit, which could be attributed to the ambiguity and complexity of Chinese characters.

Table 3 The words of top 20 topics in LDA model based on character with K=16

Topic						
1	道	段	花	两	便	中
2	道	令	狐	仙	说	师
3	手	凤	苗	赛	总	范
4	道	说	师	手	子	便
5	万	狄	丁	中	云	师
6	道	说	见	家	子	出
7	道	宝	韦	说	子	胡
8	道	靖	郭	蓉	武	七
9	道	中	手	心	身	出
10	道	中	说	子	出	心
11	国	十	回	王	三	文
12	马	镖	头	子	道	中
13	军	中	兵	官	说	宗
14	李	道	苏	秀	克	文
15	刀	道	鞭	手	招	威
16	剑	手	士	中	青	道

实际分类结果	SVM分类结果												
0	0	0	13	0	9	0	7	0	0	0	0	3	0
1	7	1	7	1	1	7	12	7	2	3	1	4	5
2	13	1	12	2	2	2	2	2	1	1	2	2	4
3	3	14	3	5	3	8	2	3	2	14	3	12	3
4	13	12	4	4	4	2	2	4	2	2	13	9	4
5	7	3	0	5	2	5	5	5	2	8	7	15	5
6	2	6	6	6	6	6	6	6	6	6	6	6	7
7	10	12	7	7	15	7	3	11	7	7	7	0	7
8	5	2	2	14	10	14	2	7	3	2	2	13	13
9	9	9	9	11	9	7	12	7	9	9	2	12	10
10	10	10	10	10	10	10	10	10	7	10	10	10	10
11	11	11	7	11	13	2	2	2	3	11	12	12	13
12	2	12	12	12	12	15	12	7	12	12	2	12	12
13	13	7	12	2	13	2	13	13	2	13	14	1	13
14	14	14	14	14	14	14	14	14	14	3	14	14	5
15	15	12	15	15	15	15	15	12	15	15	12	0	15

Fig 4 The classification result of training dataset of LDA model based on character

Conclusion

Based on the results of this report, LDA topic modeling is an effective method for classification of the Jin Yong novel corpus. The number of topics K significantly affects model construction and classification, with an overall upward trend in accuracy as K increases, but overfitting can occur when K is too large. Additionally, classification results based on " characters " as basic units are more accurate than those based on " words ". To address overfitting, the model's iteration can be reduced appropriately. The implementation of data visualization in this study aimed to use the pyLDAvis library, but some issues occurred during the process. Future improvements can be made to address these issues.