Report of Deep Learning for Natural Langauge Processing 2

Text Modeling through LDA Model

Jiayi Zhang
zy2303814@buaa.edu.cn

Abstract

This report focuses on modeling a given text using the LDA model. This paper uses Jin Yong's novels as corpus, constructs Chinese text data set by word segmentation or word segmentation, and constructs text classifier by LDA model. This paper explores the classification performance of the model under different themes, different basic unit (word, word) classification and different token quantity.

Introduction

There are a lot of random corpora in the real world, and it is difficult to extract meaningful information from them. When dealing with these irregular corpora, one feasible approach is to use Topic Models to identify the topics within documents and uncover hidden information. The topic model can be seen as a model that groups words with similar contexts in the corpus. Latent Dirichlet Allocation (LDA) is one of the most widely used algorithms for subject modeling due to its scalability and fast computation speed. It allows for generating simple, intuitive, and easy-to-understand themes.

In this study, we will employ the LDA model to predict and classify 1000 passages sampled from 16 Jin Yong novels. The classification results will undergo cross-verification ten times. We aim to compare the effects of different numbers of topics and tokens on the classification performance of the model while exploring how word-based and document-based approaches impact its effectiveness.

This paper mainly realizes and discusses the following aspects:

(1) Does the classification performance change when the number of topics T is set differently? ;

(2) What are the differences between the classification results based on words and characters?

(3) Is there any difference in the performance of the topic model between short text and long text with different number of values K?

Methodology

M1: LDA Model

LDA was proposed by Blei, David M., Ng, Andrew Y., and Jordan in 2003. It is a topic model, which can give the topic of each document in a document set in the form of a probability distribution, and then extract the topic (distribution) of some documents by analyzing them. Topic clustering or text classification can be performed based on topic (distribution). At the same time, it is a typical word bag model, that is, a document is composed of a group of words, and there is no sequential relationship between words.

Suppose there are D documents, corresponding to Ni words in the D_i document, and these documents have a total of K topics. In LDA, it is assumed that the distribution of the topic of the document is *Dirichlet distribution*, that is, for any document, the topic distribution θ_d satisfies formula (1.1) :

$$\theta_d = Dirichlet(\alpha) \tag{1.1}$$

Where α is a K -dimensional hyperparameter vector. For each topic, the prior distribution of the words therein can also be considered as a *Dirichlet distribution*, that is, for any topic k , the word distribution satisfies equation (1.2) :

$$\beta_k = Dirichlet(\eta) \tag{1.2}$$

η is a V -dimensional hyperparameter vector, where V represents the number of all the words in the vocabulary. Suppose that in document D_i, the number of the word in subject k is $n_{D_i}^{(k)}$, then the corresponding polynomial distribution can be expressed as equation (1.3) :

$$n_D = (n_{D_i}^{(1)}, n_{D_i}^{(2)}, \ldots, n_{D_i}^{(k)}) \tag{1.3}$$

The posterior distribution of θ_d can be obtained by using Dirichlet-Multi conjugation as shown in the equation (1.4):

$$Dirichlet(\theta_d | \alpha + \eta_{D_i}) \tag{1.4}$$

Similarly, we can obtain the posterior probability of β_k as follows:

$$Dirichlet(\beta_k | \eta + \eta_k) \tag{1.5}$$

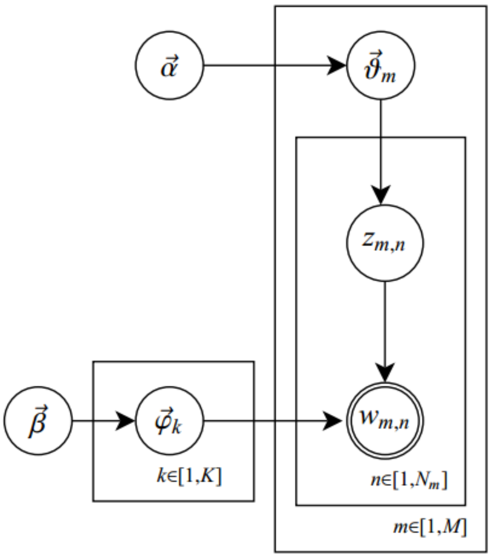

Figure 1： LDA model structure diagram

The LDA model is generally solved by *Gibbs sampling algorithm*, and its model structure is shown in **Figure 1:** The process first selects the topic number K and appropriate hyperparameters α and η, and then randomly assigns a topic number z to each word in each document in the corpus. Rescan the corpus, update its topic number using *Gibbs sampling formula*, and update the number of the word in the corpus. Then repeat these steps until the Gibbs sampling converges. The topic distribution of each word in each document in the corpus is calculated to obtain θ_d, and the topic and word distribution β_k of LDA is calculated to obtain the distribution of each subject word in the corpus.

M2:Predict article topics and K-fold cross-validation[1]

Formulas (1.1) - (1.5) were used to build the LDA model. After several iterations, the LDA model tended to be stable, and the probability of occurrence of each paragraph under the topic was calculated.

When constructing a model, it is customary to partition the data into a training set and a test set. The test set comprises data that is distinct from the training set and does not partake in the training process; rather, it serves as an evaluation tool for the final model. During training, overfitting often arises, signifying that while the model may fit the training data well, its ability to predict data beyond this scope diminishes. Utilizing test data to adjust model parameters at this stage would essentially incorporate information from some of the test data known during training, thereby impacting the accuracy of final evaluation results. A common practice involves segregating a portion of the training data as validation data to assess how effectively the model generalizes.

.Cross-validation is frequently employed for evaluating models using validation sets by dividing original datasets into K groups (k-fold), creating one verification subset per group while employing K-1 subsets as respective training sets. Consequently, K models are generated and evaluated against their corresponding validation sets before aggregating and averaging their Mean Squared Error (MSE) values to obtain cross-validation error. By making efficient use of limited dataset resources, cross-validation yields evaluation results closely aligned with performance on independent test sets—thus serving as an indicator for optimizing models.

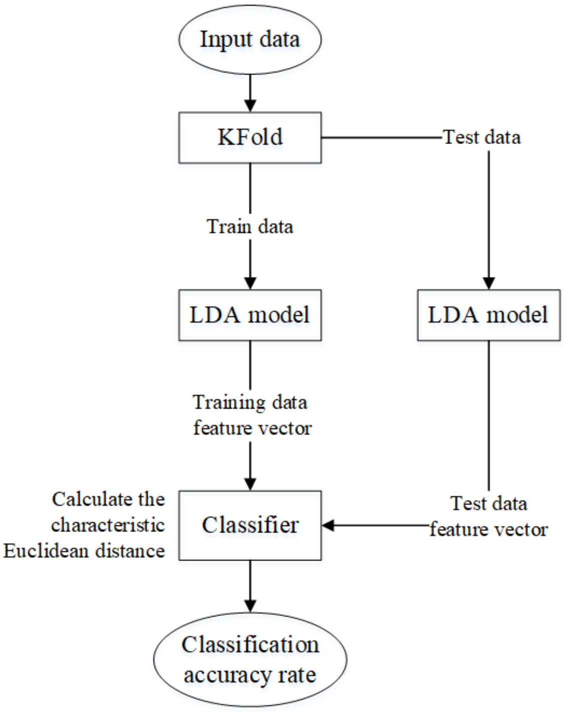

Figure 2： LDA model performance verification algorithm flow

The algorithm designed for this report is shown in **Figure 2**. Firstly, the Kfold function is used to extract 900 training sets and 100 test sets required for cross-validation from 1000 paragraphs, and LDA modeling is carried out respectively to generate topic probability distribution for each novel, namely, the feature vector of training data and the feature vector of test data. Compare whether the labels of the features that are closer to the Euclidean distance between the two vectors are the same. If they are the same, it is considered that the classification is correct and the model is successfully built. If the labels are different, the classification of the current paragraph is incorrect.

Experimental Studies

Step1: Data Preprocessing

For data preprocessing,16 novels in the jyxstxtqj_downcc.com document are read and divided into paragraphs, then use it to build a data set for classification.

(1) Go through the novel directory, read the text content of all novels one by one, save in the dictionary, each key corresponds to a novel file name.

(2) Then segmentation is carried out to generate a set of paragraphs with different feature lengths. For each novel text, it is divided into paragraphs according to the number of words of a set length K. Each paragraph is labeled with a label corresponding to the novel.

The generated data set is shown in the 10 documents "20-word-new.csv" - "3000-words-new.csv"

Step2: Modeling

Through the above experimental steps, writing and running the code, the following data can be obtained:

Table 1: Classification accuracy of **word** segmentation LDA model

Topic/Token	20	100	500	1000	3000
20	0.0784	0.2024	0.3664	0.4671	**0.6347**
50	0.0893	0.2024	0.6518	0.7671	**0.8611**
80	0.0804	0.1835	0.7319	0.8353	**0.9264**
120	0.0605	0.1915	0.7491	0.8729	**0.9486**
140	0.0724	0.1955	0.7202	0.8918	**0.9444**

Table 2: Classification accuracy of words segmentation LDA model

Topic/Token	20	100	500	1000	3000
20	0.0684	0.1088	0.1399	0.2319	**0.3490**
50	0.0526	0.0908	0.2310	0.3333	**0.4733**
80	0.0734	0.0848	0.2809	0.4395	**0.5832**
120	0.0645	0.0599	0.3050	0.4661	**0.7106**
140	0.0645	0.0649	0.3412	0.4911	**0.6838**

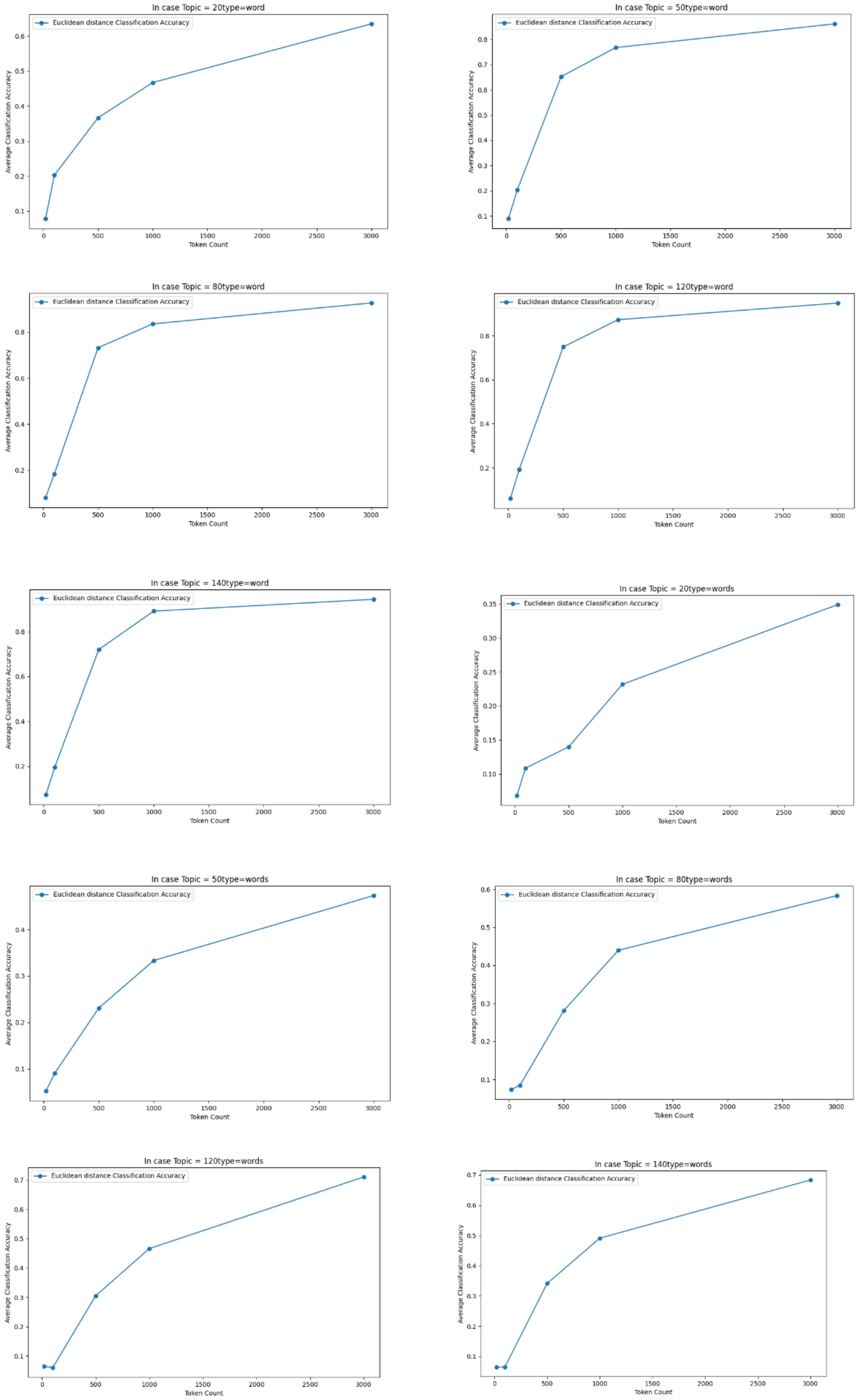

Figure 3: Experimental results of classification accuracy of LDA model

The horizontal axis of the table represents the number of tokens, while the vertical axis represents the number of topics. The token values range from 20 to 100, 500, 1000, and 3000 respectively, whereas the topic values range from 20 to 50, 80, 120, and finally reaching a maximum of 140. **Figure X** illustrates the corresponding data relationship depicted in this two table. It is evident that as the number of credit tokens increases, there is a gradual improvement in accuracy for the LDA model. Furthermore, an increase in the number of topics also leads to an observable enhancement in classification accuracy.

As can be seen from the figure, the maximum classification accuracy rate of the LDA model performance verification algorithm involved in this report can reach more than 0.9, and the data classification results based on words are better than those based on words. The analysis shows that LDA model can not understand the meaning of words, and only checks the subject corresponding to the element by word frequency. Compared with words, it is easier to find the matching relationship between individual words and words. The subjectivity of words is too strong, and the classification effect may not be ideal when the amount of data is small and the number of classified topics is small.

Conclusions

This report aims to investigate experiments that model text on a given corpus using the Latent Dirichlet Allocation (LDA) model and classify each paragraph after representing it as a topic distribution. Experiments will explore the effects of the number of topics, basic units (words and words), and text length on model performance.

Therefore, the following answers can be given to the question:

(1) When the number of topics T is set, the classification performance of the LDA model becomes better with the increase of the number of topics T;

(2) The classification effect with word as the basic unit is worse than that with word as the basic unit, which is caused by the inability of LDA model to understand semantics;

(3) For short text and long text with different values of K, the classification performance of LDA model becomes better with the increase of K.

References

[1] CSDN. KFold [EB/OL] https://blog.csdn.net/xiaohutong1991/article/details/107924703