Report of Deep Learning for Natural Langauge Processing

Xintong Zhang
zhangxintong0810@icloud.com

Abstract

The objective of this report is to perform textual feature extraction and classification on the 16 works of Jin Yong's martial arts novels. The implementation involves the following steps: First, a dataset is constructed by randomly and uniformly sampling 1,000 paragraphs from the 16 novels. The dataset is then partitioned into training and test sets using ten-fold cross-validation. Subsequently, the Latent Dirichlet Allocation (LDA) model is employed to extract latent semantic relationships, specifically documents-topics and topics-words distributions. A Support Vector Classifier (SVC) is then utilized to compute the text classification accuracy for both training and test sets. Finally, model performance curves—evaluated by character-based and word-based metrics—are plotted for two parameter configurations: variable tokens and variable topics.

The final optimal classification results are as follows: the accuracy of the training set and the test set are 75.83% and 63.37% respectively, while the accuracy of the training set and the test set are 73.51% and 68.75% respectively. Compared with the word has improved.

Introduction

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model designed to uncover latent topics from a collection of documents. Its core assumptions are: Each document is a probabilistic mixture of multiple topics (document-topic distribution)； Each topic is a probabilistic mixture of words from the vocabulary (topic-word distribution).The essence of LDA lies in reverse-engineering the generative process, i.e., inferring the latent topic structure from observed words.

The generative process can be summarized as follows: For each document d, Sample its topic distribution θ_d from a Dirichlet distribution Dir(α).

For each word $w_{d,n}$ in the document:

a. Sample a topic $z_{d,n}$ from the topic distribution θ_d .

b. Sample a word from the word distribution corresponding to topic $z_{d,n}$.

The Probabilistic Model can be described as follows： Firstly， the joint distribution of the model, given the hyperparameters

$$p(\theta, z, w|\alpha, \beta) = \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta) \right)$$

To find the marginal distribution of a document, we integrate over

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta$$

The core task in using LDA is inference, which involves determining the hidden topic structure given a set of documents. This is typically done using algorithms like:1.Gibbs Sampling: A Markov Chain Monte Carlo (MCMC) method. 2. Variational Inference: An optimization-based method that approximates the posterior distribution.

Methodology

Step1: Database loading and corpus import

First load the main functions needed such as Chinese word jieba, drawing, SVC classifier, KFlod ten cross-verification. When importing the corpus, mac computer system is incompatible with the file editing format, so the method of automatically detecting the file code is adopted to import the corpus. The import content is the text of 16 Jin Yong martial arts novels and the Chinese stop word list.

Step2: Data preprocessing:

First, the imported text is preprocessed by removing irrelevant information (for example, this book is from www.cr173.com) and removing clutter (for example, "#$%&\'()*+,-). Remove whitespace characters (such as newline \n, TAB \t '), remove related stops in the basic stop list. After the classifier function is defined, the data is read in.

Step3: Main function modeling

(1) Word segmentation:

Since the model needs to be evaluated by word and word separately, it needs to be segmented separately. For word units, we introduce the jieba library for accurate word segmentation. For word units, we traverse the text character by character, preserving Chinese characters.

(2) Uniform sampling and paragraph generation:

In order to ensure the balanced extraction of text feature information, we selected 16 novels for uniform sampling, and extracted 62 paragraphs from each, totaling about 1000 paragraphs. Label coding is used to extract the corresponding number of paragraphs and topics, and random.seed is used for random selection. If insufficient paragraphs are sampled, assert function is used to ensure consistency.

(3) Cross-verification of LDA modeling and ten fold classification:

In the data preprocessing and feature extraction phase, we first use CountVectorizer to transform the preprocessed text into a document-word matrix, where each row represents a paragraph and each column represents the frequency of a word. Then, Bayes inference method is used to estimate the distribution of each topic in the document, LDA model in scikit-learn is called, and fit_transform method is invoked to represent each document as a topic distribution vector, each component of which corresponds to a topic, and the sum of all components is 1.

After obtaining the topic distribution vector for each document, we use a support vector machine (SVM) as the classifier. In order to evaluate the classification effect, we adopted the cross-validation method to randomly divide 1000 samples into 10 different training and test combinations (900 samples for training and 100 samples for testing each time), and obtained the mean and standard deviation of classification accuracy by calculating the accuracy of each training set and test set.

Step4: Compare the model performance under different parameters and draw

In order to explore the effects of topic number and text length on classification performance and model performance, According to the control variable method to design the only consider topics change (num_topics_lst = [1,20,50,100,200,300,400,500,600]) and only consider the

paragraph length changes (paragraph_length_lst = [20, 100, 500, 1000, 3000]) The average classification accuracy of the training set and the test set under the two conditions, and the specific experimental results are analyzed as follows.

Step5: Display the result of parameter adjustment

By analyzing the classification performance of the model according to different parameter Settings in step 4, it can select the best paragraph_num=50, paragraph_length=1500, and num_topics=100, and output the final segmentation accuracy rate.

Experimental Studies

1. Does the classification performance change when the number of topics T is set differently?

The experimental results show that when the T value is low (T<50), the topic number T has a great influence on the text classification, and the model can not classify the text correctly. When T is large (T>200), the classification performance is not significantly improved, and even the classification performance of the test set is low, that is, overfitting problems are prone to occur.

To sum up, in order to ensure both generalization and model classification accuracy, and considering the impact of excessive parameters on training time and response speed, num_topics=100 was selected.

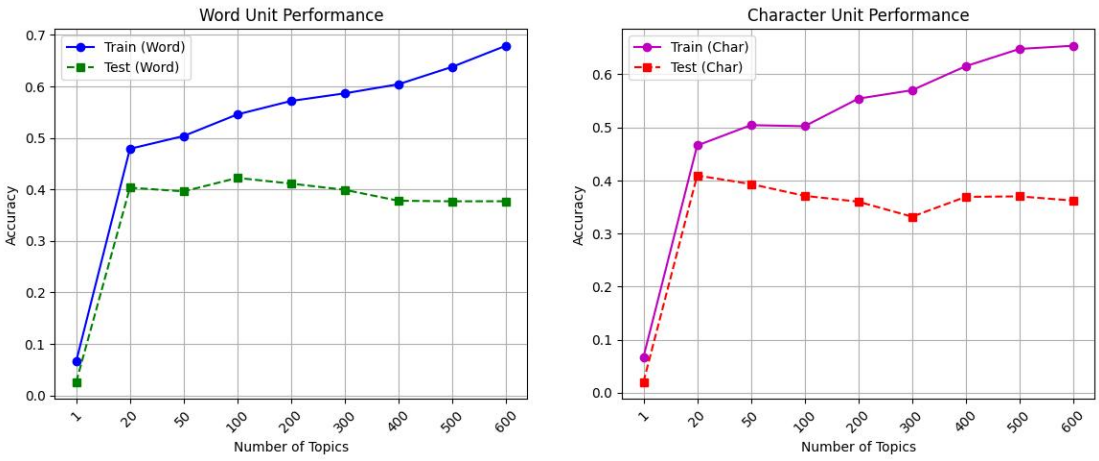

Figure 1： T arameter comparison diagram

2. Is there any difference in the performance of the topic model between short text and long text with different values of K?

According to the assignment, I compared the line plots of model classification performance when K=20, 100, 500, 1000, 3000. By analyzing the images, we can see that when K<1000, the classification performance of the model fluctuates greatly, and the LDA model may not be able to fully capture the rich semantic information in the text. When K>1000, the classification performance rises gently and stably, and the model can better depict the text theme within a certain range. This is the same as our intuitive feeling on the extraction of text feature information. The longer the text, the richer the potential semantic information we can understand, the closer the backbone information we can get and the connection between words and corresponding topics, so the longer text has greater advantages on the LDA model.

Based on the above consideration, we choose paragraph_length=1500 as the best parameter for the final parameter Paragraph_length =1500 for model performance evaluation.

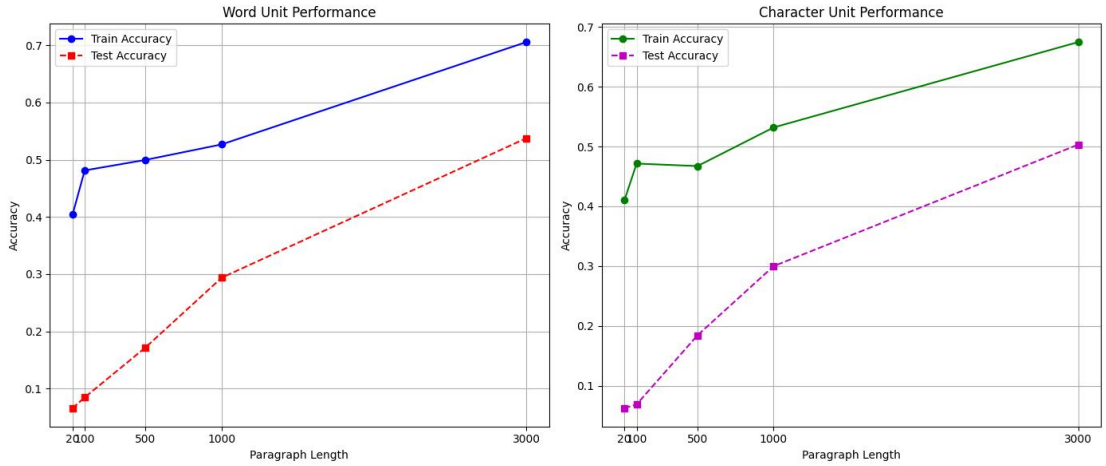

Figure 2： K parameter comparison diagram

3. What is the difference between the classification results with "word" and "word" as the basic unit?

From the above picture analysis, it can be seen that word-based LDA model has better classification performance than word-based classification, because the word-based LDA model can obtain article information more accurately and richly, thus better obtaining topic distribution and more conducive to SVM classification.

Conclusions

Finally, the best parameter paragraph_num=50, paragraph_length=1500, and num_topics= 100 are selected. The classification results are as follows: the accuracy of the training set and the test set are 75.83% and 63.37% respectively, while the accuracy of the training set and the test set are 73.51% and 68.75% respectively.

Problems and Future Improvements:

1. In view of the fact that the classification performance of some novels is always slightly higher than that of others, I personally believe that random and uniform sampling of 1000 fragments should be cancelled, and paragraph sampling should be conducted according to the proportion of the total effective characters of each text to the total effective characters of 16 novels, so as to extract the text feature information more effectively.

2. Other efficient natural language processing algorithms such as LSA and Gibbs Sampling Algorithm have not been well combined and applied, mainly because the basic connotation of the idea of "using a posterior probability to approximate the joint probability distribution" is not clear enough. I hope to further strengthen my learning in the future.

References

[1] Zenchang Qin and Lao Wang (2023)， How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.