Report of Deep Learning for Natural Langauge Processing：

Text Modeling by Latent Dirichlet Allocation

Yuqi Shi
1650278787@qq.com

Abstract

This is a Report of Deep Learning for Natural Langauge Processing about Text Modeling by Linear Discriminant Allocation.This report will use the LDA model to perform text modeling on a given corpus with a topic count of T, where each paragraph is represented as a distribution over topics, followed by classification,and solve the following three problems.Firstly,Does classification performance vary with different numbers of topics T set in the LDA model? Secondly,What differences arise in classification outcomes when using "words" versus "characters" as the fundamental units?Thirdly,does the performance of topic models differ for short and long texts with varying values of K (number of topics)?

Introduction

LDA (Latent Dirichlet Allocation) is a probabilistic graphical model-based approach for text topic analysis. Initially proposed by Blei and colleagues in 2003, it aims to automatically uncover hidden thematic structures within textual data through analysis. The core idea behind the LDA model is to represent texts as a set of probability distributions, wherein each document is a mixture of multiple topics, and each topic, in turn, consists of a collection of words

The fundamental principle of the LDA model involves postulating a generative process for a collection of texts. Specifically, it assumes the following steps: First, a topic is randomly selected from a topic distribution. Next, a word is then chosen at random from the word distribution associated with that topic. This procedure is repeated, generating words one by one, until an entire document is formed. This process is iteratively applied to create the entire collection of texts, thereby capturing the underlying theme composition of each document.

Specifically, the generative process of the LDA model comprises three steps:

1.Selecting the document's topic distribution: Randomly choose a topic distribution from the Dirichlet Distribution

2.Selecting the topic for the document: For each position in the document, randomly select a topic from the topic distribution.

3.Selecting the word: For each position in the document, randomly pick a word from the word distribution of the chosen topic.

Methodology

The research will be introduced in 4 parts :text preprocessing, LDA model construction,train and evaluate the LDA model ,and results visualization.

M1:Text Preprocessing

Before conducting topic modeling, it is necessary to preprocess the text, which includes tasks such as tokenization, removing stop words and punctuation, among others. Tokenization can be performed using tools like Jieba in the case of Chinese text.

M2: LDA Model Construction

1. **Build the dataset**:From the given corpus, uniformly sample 1,000 paragraphs to form a dataset. Set each paragraph to have K tokens, where K is set to 20, 100, 500, 1000, and 3000 respectively. The label for each paragraph is the novel to which it belongs. Conduct experiments using both characters and words as the unit of a token.
2. **Construct dictionary and documents:**LDA employs the Bag-of-Words (BoW) model, which, in essence, represents a document by considering only whether a word occurs within it, disregarding the order in which words appear.In addition,using the Gensim library, one can create a dictionary of documents and a document-term frequency matrix. The dictionary encompasses all the words present across all documents, while the document-term frequency matrix denotes the frequency of each term in every individual document.Vary the number of topics T to 8, 16, 32, and 64, and employ the `gensim.models.LdaMulticore` function, inputting the paragraph data and the specified number of topics, to construct the LDA models.
3. **Train and evaluate the LDA model:**Utilizing the LDA model class from Gensim, one can execute the LDA algorithm,to train and evaluate the model.
4. **Results visualization**:After representing each paragraph as a topic distribution, classify them using a Naive Bayes classifier constructed with the `MultinomialNB()` function from the `scikit-learn` module. The classification results are evaluated using 10-fold cross-validation to obtain the average accuracy.

Experimental Studies

Table 1:Classification accuracy

	Topic								
	以字为单位				以词为单位				
K	8	16	32	64	k	8	16	32	64
20	0.177	0.124	0.157	0.126	20	0.132	0.164	0.147	0.132
100	0.141	0.192	0.166	0.146	100	0.247	0.317	0.328	0.276
500	0.237	0.255	0.339	0.395	500	0.518	0.626	0.653	0.628
1000	0.381	0.427	0.509	0.543	1000	0.594	0.772	0.778	0.749
3000	0.449	0.646	0.714	0.735	3000	0.787	0.877	0.879	0.89

As we can see from Table 1,when the paragraph length K remains constant, increasing the number of topics T leads to higher accuracy. Conversely, when the number of topics T is held constant, longer paragraphs, represented by larger values of K in the case of short and long texts, yield higher accuracy rates.Moreover,using "words" as the basic unit yields much higher accuracy and better classification performance compared to using "characters" as the fundamental unit.

Conclusions

A Chinese corpus which includes 16 novels of Jin Yong is used as the data to verify the text classification performance by LDA model.Through the experiment,I conclude that changing the paragraph length,the numbers of topics and the basic unit can improve the classification performance.With an increase in the number of topics and the length of paragraphs, adopting "words" as the basic unit leads to even higher accuracy and superior classification effectiveness.

References

[1] https://zhuanlan.zhihu.com/p/31470216

[2] https://blog.csdn.net//qq_41667743/article/details/129378418