Paragraph Classification of Jin Yong Novels Based on Topic Modeling

Hao Lu

lh2002@buaa.edu.cn

Abstract

This experiment investigates the use of topic modeling (LDA) for short text classification. Using a corpus of 16 martial arts novels by Jin Yong, we randomly sample 1,000 paragraphs, each labeled by its source novel. By varying the number of topics T, the textual unit (character vs word), and paragraph length K, we analyze their impact on classification performance.

Introduction

The experiment is designed to explore the following questions: Does the number of topics T affect classification performance? How does the choice of basic unit ("word" vs "character") influence classification results? How does paragraph length K (i.e., short vs long texts) affect the effectiveness of topic modeling?

Methodology

Data Preprocess

Data source: 16 Jin Yong novels in .txt format.
Sampling: Randomly and evenly sample 1,000 paragraphs from all novels, each containing at least K tokens.
Text unit: Paragraphs are processed at either the word or character level.
Label: Each paragraph is labeled with its corresponding novel.

Params

Paragraph length K: 20, 100, 500, 1000, 3000
Text unit: 'character', 'word'
Number of topics T: 10, 20, 50, 100, 200
Classifier: Logistic Regression

Evaluation method: 10-fold cross-validation (900 training, 100 testing)

Method Details

Text data is first tokenized using Jieba for word segmentation. Paragraphs containing fewer than K tokens are discarded to ensure sufficient context for topic modeling. Each remaining paragraph is then stored along with its corresponding label (i.e., the novel it originates from).
Topic modeling is performed using Gensim's LdaModel, where each paragraph is converted into a T-dimensional topic distribution vector, representing the paragraph's probability over the T learned topics.
Each topic distribution vector is used as input to a Logistic Regression classifier implemented with scikit-learn. The model's performance is evaluated using 10-fold cross-validation, and classification accuracy is reported as the main evaluation metric.

Experimental Studies

Varying the value of T

Regardless of whether characters or words are used as the basic unit of tokenization, the performance of the LDA model improves significantly as the number of topics T T increases, reflecting the benefits of richer semantic features. However, when T T becomes too large, performance may degrade, possibly due to overly sparse topic distributions that make learning more difficult. Based on the experimental results, a topic number around 50 appears to be the most appropriate for this task.

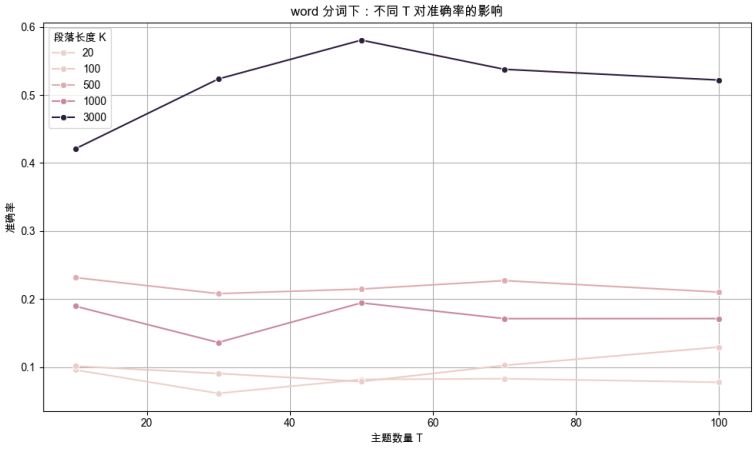

Figure 1: T trending with word split

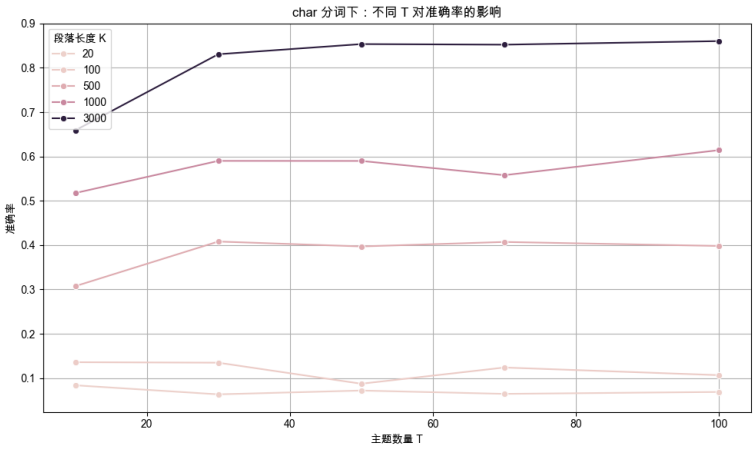

Figure 2: T trending with char split

Word v.s. Char

In general, words are considered to carry richer semantic information than characters, and thus using words as the basic tokenization unit is typically expected to yield better performance. However, in this experiment, using characters as the unit of segmentation resulted in significantly better performance. A possible explanation lies in the nature of the Chinese language: individual characters in Chinese often already contain sufficient semantic information. This is supported by the entropy analysis in the first experiment. Given the limited size of the corpus in this task, using overly informative units such as words may fail to capture statistically reliable patterns, whereas character-level representations strike a better balance between granularity and statistical robustness.

Table 1: LDA Classification Accuracy (unit=char)

T	10	30	50	70	100
K					
20	0.0838	0.0634	0.0720	0.0645	0.0688
100	0.1359	0.1348	0.0873	0.1239	0.1067
500	0.3074	0.4081	0.3971	0.4072	0.3981
1000	0.5175	0.5901	0.5900	0.5578	0.6143
3000	0.6589	0.8307	0.8536	0.8523	0.8603

Table 2: LDA Classification Accuracy (unit=word)

T K	10	30	50	70	100
20	0.0958	0.0613	0.0818	0.0829	0.0776
100	0.1013	0.0905	0.0787	0.1024	0.1294
500	0.2315	0.2079	0.2147	0.2270	0.2102
1000	0.1895	0.1361	0.1943	0.1712	0.1712
3000	0.4208	0.5235	0.5807	0.5377	0.5220

Varying the value of K

As K K increases, the performance of the LDA model improves significantly. This is because longer paragraphs contain more robust and stable features, leading to better topic modeling and classification results.

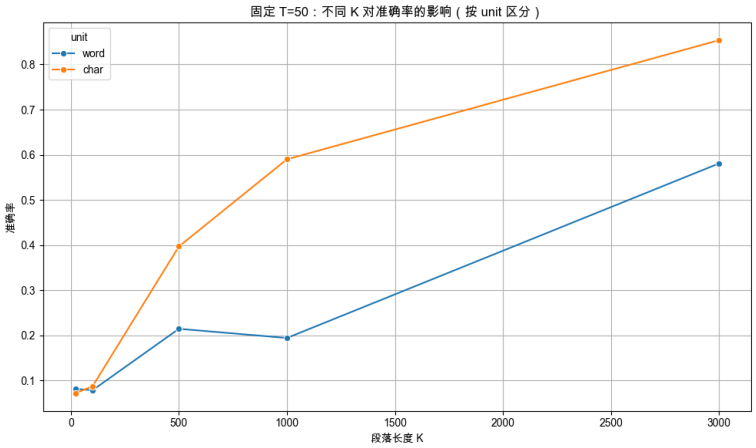

Figure 3: Fix T=50

Conclusions

In this experiment, we explored the impact of different factors—tokenization units (character vs. word), paragraph lengths (K), and topic numbers (T)—on the performance of LDA-based text classification. The results show that increasing the number of topics generally improves performance up to a certain point, with optimal results typically occurring around T=50. Interestingly, character-level tokenization consistently outperformed word-level tokenization, which may be attributed to the semantic richness of individual Chinese characters and the limited size of the dataset. Additionally, increasing

the paragraph length K significantly enhanced performance, as longer texts provide more robust and reliable statistical features for topic modeling. These findings highlight the importance of careful parameter selection in topic modeling tasks, especially when working with Chinese text and limited data.