Report of Deep Learning for Natural Language Processing

MengCao ZY2403801

buaa_cm@buaa.edu.cn

Abstract

This paper uses the corpus of Jin Yong's novel《倚天屠龙记》 to train word vectors based on Word2Vec model, and visually tests the validity of testimony vectors through word similarity calculation and word vector clustering. The experiment shows that Word2Vec model can capture the semantic correlation of characters, martial arts, weapons and other words in Jin Yong's novels, and provide effective support for the semantic analysis of martial arts texts.

Introduction

In the field of natural language processing, word vector training is an important basis for understanding text semantics. As a classic of wuxia literature, Jin Yong's novels contain rich vocabularies such as characters, martial arts and school. This study aims to use Word2Vec model to train the word vector of Jin Yong's novel corpus, verify the model's ability to capture semantic relations of Wuxia texts, and provide a basis for subsequent wuxia text analysis (such as character relationship mining, plot clustering, etc.).• The influence of subject number T on classification accuracy;

Methodology

Word2Vec

Word2Vec is a neural network based word vector training model, the core includes continuous word bag model (CBOW) and Skip word model (skip-gram). The Skip-Gram model is adopted in this experiment, and its principle is to predict the context words with the current word as the center. The model learns the distributed representation of words by optimizing the

probability of "generating context words according to the central word", so that semantically similar words are closer together in vector space. In the training process, negative sampling and other techniques are used to improve the efficiency and finally generate low-dimensional dense word vectors to represent lexical semantics.

Experimental method

Data preprocessing: cleaning the text of Jin Yong's novels, using stuttering word segmentation after sentence segmentation, combining with the Wuxia dictionary to improve the accuracy of word segmentation, filtering stop words, single words and short sentences.

Word vector training: Using Word2Vec model, set vector dimension 200, window size 8, minimum word frequency 10 and other parameters to get the word vector training.

Validity verification:Word similarity calculation: The cosine similarity of words such as people and clans is calculated to verify the model's ability to capture semantic associations.

Clustering visualization: PCA dimension reduction is used to visualize the word vector clustering of characters, martial arts and weapons words, and to observe the spatial clustering of words with similar semantics.

Experimental Studies

Word similarity analysis

The similarity of multiple groups of words is calculated in the experiment, and the results are as follows:

Character relationship: The similarity of "张无忌 - 赵敏" is 0.4821, which is higher than that of "张无忌 - 周芷若", reflecting that the model captures a closer relationship（最后成为夫妻） between 张无忌 and 赵敏 in the novel, and in fact, 张无忌 chooses 赵敏 as her partner in the novel. The similarity of the same 灭绝师太 as 周芷若's teacher is 0.3050, higher than that of 赵敏's 0.2059.

The similarity of "周芷若-峨嵋派" is 0.2257, which is slightly higher than that of 张无忌 and 峨嵋派 (0.2192), indicating that 周芷若-峨嵋派 is more closely related to each school, and it is reasonable that 张无忌, as the protagonist, is related to major denominations;

The similarity between "周芷若 and 明教" is only 0.0755, which accords with the setting that 周芷若 and 明教 are weakly related in the novel.

Table I Word similarity analysis

词语对		相似度
张无忌	周芷若	0.3963
张无忌	赵敏	0.4821
灭绝师太	周芷若	0.3050
灭绝师太	赵敏	0.2059
张无忌	明教	0.2953
张无忌	峨嵋派	0.2192
周芷若	明教	0.2257
周芷若	峨嵋派	0.0755

Cluster visual analysis

The dimensionality of the word vector is reduced to two dimensions by PCA and visualized (see figure).

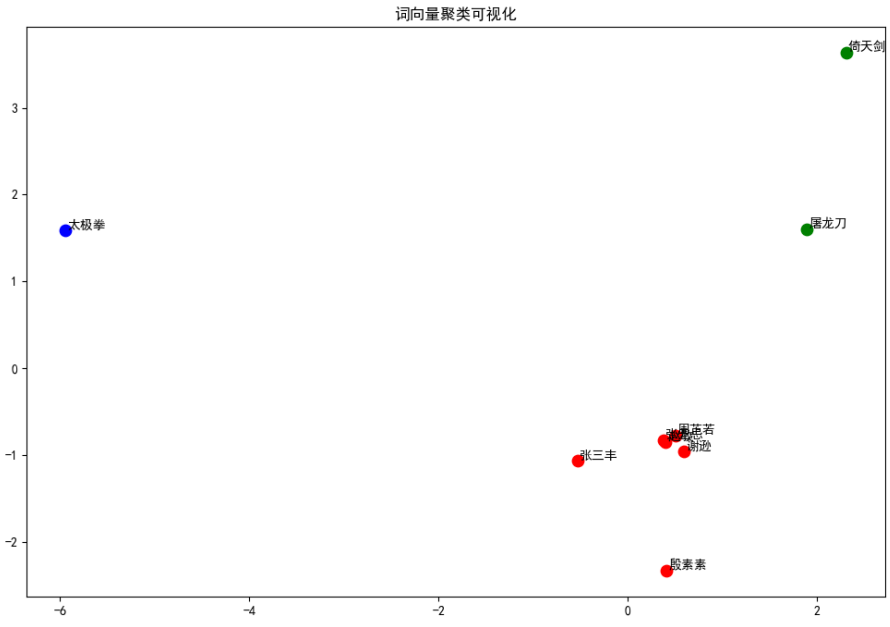

Figure 1 Cluster visual analysis

Characters (red) such as "张无忌" and "周芷若" gather, reflecting the semantic correlation of characters;

Weapons (green) "倚天剑" and "屠龙刀" are similar in location, reflecting the semantic aggregation of similar items;

In the martial arts category (blue), only “太极拳” is shown, but it can be seen that it is far away from characters and weapons, and the rationality of word vector construction can be seen.

Conclusion

In this study, the word vector of Jin Yong's novel corpus is trained by Word2Vec model, and its validity is verified by similarity calculation and clustering visualization. Experiments show that the model can effectively capture lexical semantic associations, such as the relationship between people and the affiliation of clans. In the future, we can optimize the corpus size, try more complex models (such as GloVe), further improve the quality of word vectors, and expand to text generation, plot analysis and other application scenarios.

References

[1] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[2] 何晗。自然语言处理入门 [M]. 人民邮电出版社，2019.

[3] Jieba Chinese Text Segmentation: https://github.com/fxsjy/jieba