REPORT

Yu Feng z5094935

November 2017

1 Introduction

In this assignment, we implement a word embedding project by tensorflow.

1.1 Dataset

the data set is from BBC news. There are about 0.7 million words from BBC news. It is a retentively small data set.

2 Introduction

There are two major model for Word2Vec – skip-gram and CBOW. The skip-gram model perform better for lager date set. So I choose CBOW model.

2.0.1 pre-process

By Spacy I transfer some words to it's class such as location , money, name and so on. And cleaning the dataset by get rid of space and punctuation, removing the words only occur once.

2.1 parameter tuning

After grid search, I found that 10 negative sample, 3 skip window, default learning rate is best choose.

2.2 model evaluate

At first, we do not have ground true. It's really hard to judge which result is better. So I choose the best manually. And look up the loss value, stop training when it converge to avid overfitting.
after release ground true. I calculate the F1 score.

2.2.1 Optimize

For calculate top k function, the program only output the word which has the same position compared with query word.
I tried soft skip window - the words have more possibilities to be selected if it's near the centre word, but is does not performance well. and I tried get batch by probabilities - the word occurred more often has less chance to be selected. It still performed worse than original one.

3 Conclusion

Compared with ground true. The percision for top 10 word is about 10%.