COMP 6714 Project2 Word Embedding Report

Z5154763 Yihao WU

1. Introduction

The aim of this project is to generate the word embeddings for adjectives which means adjectives that have close distance in word embeddings are synonym relationship. The similarity between two words is computed using the cosine similarity measure.

2. Process tokenization

2.1 Punctuation and number removal

Using the isalpha() method to judge if it is a word or not. If it is a number or punctuation, it will be removed.

2.2 Case folding

Change all letters to lowercase by using lower() method.

3. Experimental process

3.1 Training method and data

Data set: BBC_Data.zip
Word2Vec model: skip-gram
Language process tool: spaCy
NLP process: Punctuation and number removal, Case folding

3.2 Parameters

Batch_size: 64
Skip_window: 2
Num_samples:4
Vocabulary_size: 15000
Learning_rate: 0.003
Number_of_negative_sample: 800
Embedding_dimensions: 200-dimensions
Number_of_iterations: 100001
Loss function: sampled_softmax_loss
Optimization Method: AdamOptimizer

3.3 Batch generation

After several experiments, skip_window = 2 and num_samples = 4 is chosen as the batch parameters in this project.

3.4 Hyperparameters selection

There are three hyperparameters we need to choice for better performance. They are learning rate, vocabulary size and number of negative samples. AdamOptimizer is an optimizer which need a small learning rate, so we test the learning rate in the range of 0.001-0.005.

Learning rate	Average hit
0.001	4.02
0.002	4.13
0.003	4.42
0.004	4.25
0.005	4.01

As shown in the table, 0.003 is the best choice, so I choose learning rate 0.003 for AdamOptimizer in this project.

Then we choose vocabulary size and number of negative sample

Vocabulary size	Number of negative samples	Average hits
15000	1000	8.14
15000	800	8.23
15000	600	7.63
15000	400	7.27
15000	200	7.04

Vocabulary size	Number of negative samples	Average hits
15000	800	8.79
12000	800	8.41
10000	800	8.37
7500	800	8.36
5000	800	8.20
4500	800	8.02

To sum up, the hyperparameters can be chosen as:
Batch_size: 64
Skip_window: 2
Num_samples: 4
Vocabulary_size: 15000
Learning_rate: 0.003
Number_of_negative_sample: 800

4. Conclusion

Training Word Embeddings with NLP process (Punctuation and number removal, Case folding) and an Improved version of skip-gram can be very effective to implement Word Embeddings for adjectives obtaining embeddings to preserve as much synonym relationship as possible.