COMP6714 17s2 Project 2

z5121987 Chenxuan Rong

November 16, 2017

1 Design and Implementation

This project implemented Word Embeddings for adjectives which can obtain embeddings to preserve as much synonym relationship as possible. The raw data set BBC Data.zip is used to generate the word embeddings. By using spaCy, raw data has been processed to get rid of a couple of entities, punctuationnumbers and only with 15000 commonest words. During several experimental training with using lemmatization and replace entities, the average hits has been improved.

1.1 Optimization

(a) During proprocessing stage, I removed puncuation and use entity recognition to replace those words with 'ENTITY'

(b) Using dependency parsing function from SpaCy to detect phrase word and store them in the data list.

1.2 Method to find topK

Two approaches were tested during the stage.

(a) Only writting adjective into final embedding, the drawback is the embeddings vocabulary size would be relatively small.

(b) Writing all words into embedding while during returning most similiar words stage, only adjective are returned.

1.3 Possible Improvements and Extensions

Batch execution technique could be used to improve efficiency in terms of tuning hyper parameters,

2 Attributes

To the best of my knowledge, all submitted code is my own.