COMP6714 Project2 Report

1.Introduction

In this project, we are going to implement our own Wording Embedding for adjectives. Moreover, we should obtain embedding to preserve as much synonym relationship as possible.

2.Methodology

At the beginning, we should process the data from the 'BBC_Data.zip'. This zip file contains several folders and for each folder there are many text files. Now the first thing we should do is that we need to read all these files line by line. At the same time, we are using **spaCy** to process the words in each line. We could know the characteristic of each word by using the spaCy function named '**token.pos_**' and I use this function to remove the numbers and punctuations, which could influence the results of training. Moreover, I try some methods to preprocess data before using sapCy. I try to lower case all the words and use decode to transfer the files into the 'utf-8' pattern, so that it becomes friendly to read. Then after using spaCy, I preserve the processed data into a file named 'processed_data.txt'.

In the next step, we are using **word2vec skip-gram model** to train the embedding of the data we processed in the previous step. I try some different numbers about the parameters and I decided to use the learning rate equals to 0.002 and the vocabulary size is 15000.

Then we preserve the content of **final_embeddings** and **reverse_dictionary** into a text file named 'adjective_embeddings.txt', which could be loaded by Python library genism.

Finally, we could use genism to get the results of **top_k** most similar synonym adjective words.

3.Results and Discussion

The final results about average hit is nearly 7, I think the step of preprocessing could do more things like stemming or lemmatization so that the final results could be higher.

4.Summary

By doing this project, I have learned some basic ideas about data preprocessing and word embedding. Because the training time is long in this project and the time limit, I think the final results still could be improved.