COMP6714 Project 2 Report

Z5098663 Yichen Zhu

i) Abstract/Summary

The project constitutes two major parts: Data Processing and Training Data Generation and Model Training.
In data processing, I used regex to filter some words like currency and some percentage figure as its really hard to normalize. Figure cant reflect a common sense. And I also filter some figure and word combing word. Then I use Spacy to process with 5 strategies: 1. filter non-alpha token 2. punctuation and blank 3. transfer all letter to lower case 4. if token VERB, get its lemma 5. if token is an entity, store as an entity.
In model training, I implement skip-gram model with skip window = 1 to generate training data between a center word and a random sample of its context.

ii) Introduction

The project constitutes data processing and model.
Numpy, tensorflow, spacy, gensim and re libraries are imported in this project.
Numpy is to store data.
Tensorflow is to build, train and test model.
Gensim is to store and load the model.
Skip-gram model has below parameters:
1.batch_size 128
2.skip_window 1
3.num_samples 2
4.Vocabulary_size 20000
5.learning_rate for optimizer 0.001
6.Number of Negative Samples 64

iii) Methodology

Preprocessing is the main workload for this project.I used regex to filter some words like currency and some percentage figure as its really hard to normalize. Figure cant reflect a common sense. And I also filter some figure and word combing word. Then I use Spacy to process with 5 strategies: 1. filter non-alpha token 2. punctuation and blank 3. transfer all letter to lower case 4. if token VERB, get its lemma 5. if token is an entity, store as an entity.
And from the result we can see the processing is necessary because it eliminates some words with little effect on other.

iv) Results and Discussion

After 100001 iterations, the model can hit around 7 with the current dataset.
Because we assume spacy produce right answer so some domain knowledge are

still needed to cope with this problem. And in this model, data processing is still the most important part.

v) Conclusion

Skip gram model can predict similar adj but not so efficient. Also, better processing is needed with domain knowledge.