TDDE09 Lab 5X report – johmy592 erino397

Problem 1:
When trying ourselves we get an accuracy of about 0.8.

Problem 2:
We tried training our embedding from lab5 on the oanc data. We discovered that the vocabulary missed many of the words in the toefl.txt file with threshold=100 in the make_vocab() method, so we tried to lower the threshold to 25 in order to include more words. The results were as follows:

Accuracy with threshold=100 (vocab size = 9 313): 0.425
Accuracy with threshhold=25 (vocab size = 22 577): 0.5125

For reference the vocab size when using the word2vec embedding on the same data set is 53 923.

Problem 3:
It takes about 90 seconds to train the word2vec embedding on the oanc data.

Problem 4:
We get an accuracy of 0.61.

Problem 5:
Training the word2vec embedding on the wikki data takes around 10 minutes and gives an accuracy of 0.8.

So it takes 6.67 times longer to train on the wikki data and the accuracy is 1.33 times better.
The oanc data contains 11.4 million words while the wikki data contains 124.3 million words.