COMP814 – Text Mining Lab

Skip-gram Lab

Objective

1. To modify an existing Skip-gram model and run it with different parameters

Task

1. You will need the sample code on skip-gram from the lecture.
2. Study the various part of the code in pairs and ensure that you

   understand what they do.
3. Comment out the part of the code that does the visualisation.
4. Add in code to compute the distance of each of the words to each of the other words. You can use the dot product to calculate the cosine distance between each of the words.
5. Output 2 nearest neighbours (context words) for each of the words.
6. Change the number of embeddings in the dimensions to 4 and now calculate the distance for each of the words to the other words.
7. Again find the 2 nearest context words for each of the words.

8.Additional Challenge for your own learning

9. Implement the CBOW model. That is, given 2 context words, find the nearest word to it.
10. Obtain a medium size corpus (say about 2000 documents) from online sources.
11. Find the 4 context words for 10 randomly chosen words.
12. Find the focus word for 10 sets of 4 context words.
13. Comment on whether the results seem accurate or not from your own knowledge of the meaning of the words.