University of Lethbridge

Department of Mathematics & Computer Science

CPSC 4310/5310/7310 – Spring 2020

Natural Language Processing (NLP)

Assignment 3 [75 points]

Due on March 20th, 2020

The problems are adopted from the textbook.

1. **[10 points]**

   Given the following word-context matrix:

	aardvark	computer	data	pinch	result	sugar
apricot	0	0	0	1	0	1
pineapple	0	0	0	1	0	1
digital	0	2	1	0	1	0
information	0	1	6	0	4	0

   Give the cosine similarity between all the possible pairs of words.

2. **[10 points]**

   Given the following documents:

Document 1	food restaurant customer restaurant waitress
Document 2	food store customer cashier
Document 3	appliance store customer store cashier

   Using the TF-IDF, show how similar is each pair of documents?

3. **[10 points]**

   Find tagging errors in each of the following sentences that are tagged with the Penn Treebank tagset:

   (a) I/PRP booked/VB a/DT flight/NNP from/IN Lethbridge/NNP to/TO Calgary/NNP ./.

   (b) Does/VBZ this/DT flight/NNP serve/VB complementary/JJ drinks/NNS ?/?

   (c) I/PRP have/VBP a/DT friend/NN living/VB in/IN Calgary/NNP ./.

   (d) Can/MD you/PRP list/VB the/DT afternoon/RB flights/NNPS ?/?

4. **[15 points]**

   Use the Penn Treebank tagset to tag each word in the following sentences:

   (a) It is a sweet dream.

   (b) The new store is close to the restaurant on the 13th street.

   (c) Give it a quick thought when you have a spare time to kill.

5. **[30 points]**

   Given 2 text documents, write a function that computes the similarity between the 2 documents using the cosine similarity measure with:

   (a) TF-IDF representation

   (b) Word2Vec representation

   Apply your function to the Brown corpus to compute the similarity within the cluster and between clusters.