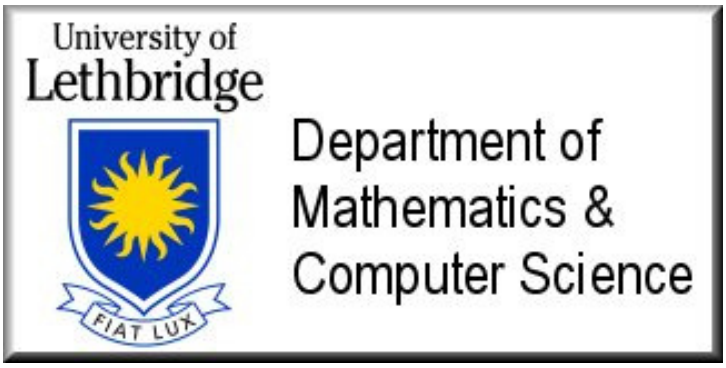

CPSC 4310/5310/7310 – Spring 2020

Natural Language Processing (NLP)

Assignment 2 [60 points]

Due on February 7th, 2020

The problems are adopted from the textbook.

1. **[5 points]**

   Write out the equation for trigram probabilty estimation in the language model.

2. **[10 points]**

   Using the naive Bayes classifier, show how the following unseen example would be classified?

	Doc	Words	Class
Training	1	Flames Oilers Senators	h
	2	Penguins Flames Predators	h
	3	Flames Oilers Canucks	h
	4	Flames Raptors Canadians	h
	5	Raptors Cavaliers Knicks	b
	6	Raptors Canadians Spurs	b
Testing	7	Flames Penguins Oilers	?
	8	Raptors Oilers	?
	9	Flames Penguins Raptors Oilers	?

3. **[15 points]**

   Using a language model generated from the Brown corpus, write a program that generates random sentences (i.e., Shannon Visualization Method) using:

   (a) bi-grams

   (b) tri-grams

4. **[30 points]**

   Using the Brown corpus, write a program that classify texts according to one of the categories/genres, and evaluate your program. You should use:

   (a) Naive Bayes

   You should split the data into training and testing (e.g., 70% and 30%), mix up the testing data and see if your classifier will succeed to split them into categories/genres again.