**ITCS 4111/5111 Introduction to Natural Language Processing**

**Assignment 4 Due: October 27th, 2017 11:59 pm**

The problems marked with \*\* are optional for students enrolled in the ITCS 4111 section of this course, all other problems are mandatory. ALL problems are mandatory for the students enrolled in ITCS 5111.

**Problem 1 (Text Classification) (50 points).**

In this assignment, you will build various text classification models and use them to classify sentences from 2016 presidential debates according to speaker.

The Assignment 4 zip file contains three files train, dev and test. The files are one document per line, tokenized and lowercased, so you don't have to do any preprocessing. Each line has the format:

trump i fully understand .

where the first token is the speaker's name, and the remaining tokens are the words of the document.

Implement a Naïve Bayes classifier using python code. Write code to read in the training documents and collect counts of documents per class and counts of words for all speakers. Write code to compute the probabilities. Use log probabilities to avoid underflow issues. Run the classifier on the test set and report your accuracy, which should be at least 50%. Describe any implementation choices you had to make (e.g., smoothing, log-probabilities).

**Problem 2 (Feature Engineering) (30 points).**

Experiment with (at least) two new kinds of features.

1. Extend the code you wrote to construct a bag of words to include an additional kind of feature besides words. You can try bigrams, prefixes, parts of speech, anything you like. Describe your new features. Report your new accuracy. Briefly write down your conclusions from this experiment.
2. Do the same thing for another kind of feature.

**Problem 3 (Feature Validation)\*\* (20 points).**

Plot the 20 most representative words in each speaker’s vocabulary sorted by their weight according to your model using R.

Note: ITCS 5111 students will be graded out of a 100 points total for all problems in this assignment. ITCS 4111 students will be graded out of 80 points total for all problems in this assignment, except for those marked with \*\*.