

# ENGS/QBS 108 Fall 2017 Assignment 4 Part 1

Due October 31, 2017

Instructors: George Cybenko and Saeed Hassanpour

Prepared by: Benjamin Priest

**Problem: Natural Language Processing [50 points].** In this problem, you will exercise your knowledge of natural language processing by way of a few problems. Further problems concerning reinforcement learning will be assigned later in the week.

1. [20 points] We will begin by building a language model utilizing [War and Peace](#).
  - (a) Change all alphabetic characters to lowercase and all non-alphabetic characters in the text to new-lines (If you have time, read the novel first, before changing it).
  - (b) Tokenize the text file and output the first 10 words of the book.
  - (c) Sort the words alphabetically and output the first 10 words of the book.
  - (d) Count the unique words in the book and output the first 10 (unique) words in the alphabetical order and their counts.
  - (e) What are the 10 most frequent words in the book and what are their frequencies?
2. [15 points] Calculate the minimum edits distance between these pairs of [Barbapapa family](#) members' names. Assume the penalty for a gap is 1 and for a mismatch is 2.
  - (a) barbapapa and barbamama
  - (b) barbabravo and barbabright
  - (c) barbabeau and barbabelle
  - (d) barbalala and barbalib
3. [15 points] Assume we have the following documents in the "Doc Label: Doc Text" format (Prof. Hassanpour will cover this topic in the next class):
  - NH: Hampton, Hanover, Keene, Concord, Manchester
  - NH: Dartmouth, Manchester, Hanover
  - NH: Manchester, Hanover
  - VT: Middlebury, Burlington
  - VT: Stowe, Rutland, Middlebury, Burlington

Use this training set and naïve Bayes text classification framework with Laplace smoothing to predict the labels for the following documents:

- (a) Dartmouth, Hanover, Burlington
- (b) Middlebury, Dartmouth, Manchester, Burlington
- (c) Stowe, Manchester, Manchester, Burlington
- (d) Keene, Rutland, Stowe