Distributed Computing for Data Science- Assignment #1
Due: Tuesday, February 23rd

This assignment is intended to provide students with increased familiarity with writing map-reduce algorithms.

You will be running your programs on a toy example (the first 10 chapters of a book by my favorite author) using a standalone (not-distributed) installation of Hadoop.

I have made you each an account on hadoop1 (10.10.11.67). Your username is your first name. Your initial password was given to you during class.

I am in the process of making more machines available with a hadoop installation. I will post details on the Moodle in the near future.

You can also find on Moodle the datafile for use in this assignment.

Problem # 1:

Write a map-reduce program to count the total number of each of the 5 vowels (A, E, I, O, U) in the data file.

Problem #2:

Part A) What word most often ends a sentence? (That is, appears immediately before a period, question mark or exclamation point)

Part B) How many different words appear at the end of a sentence?

Problem #3:

Part A) Which word occurs most frequently in the datafile?

Part B) Which word most commonly follows the word "the"? (You may ignore occurrences of "the" at the very end of a line.)

Part C) Which word has the largest number of distinct words that follow it?

Problem #4:

Part A) Write a map-reduce program to determine the average (mean) number of vowels in a word.

Part B) Modify your map-reduce program to also determine the average (mean) number of vowels in a line. That is, create a single map-reduce job that outputs both averages (vowels per word and vowels per line).

Problem #5:

Part A) Implement a Linear Regression to model word length as a linear function of number of vowels. (<Word Length> = A * <Number of Vowels> + B)

Part B) Calculate the average squared residual for your model. (That is, what is the average squared deviation of the data from your model.)

Part C) Consider an alternative model that hypothesizes that Word Length is a linear function of the square root of the number of vowels. (<Word Length> = A * <Sq Root of Number of Vowels> + B)

Part D) Compare the average squared residual for the two models.