CS246H: Mining Massive Datasets Hadoop Lab **Winter 2014**

Problem Set 3

Due 12:30pm Frebruary 5, 2014

General Instructions

As with the remainder of homework's in CS246H (but **NOT** in CS246!), this homework is optional and intended to give you practice with Hadoop. The focus of this assignment is just to practice working with MapReduce. For all problems in this assignment, use the *Complete Works of William Shakespeare* from Project Gutenberg at http://www.gutenberg.org/cache/epub/100/pg100.txt as the input dataset.

Questions

1. Use Hadoop streaming to recreate the WordCount application without writing any code.

   Hint: You can replace all non-word characters with newlines on the command line with:

   ```
   cat input | tr -sc 'A-Za-z0-9' '\n'
   ```

   You can also count the number of consecutive lines that are the same with:

   ```
   cat input | uniq -c
   ```

   For this question, please place the command-line used into a file (script) and upload the script.

2. Use Hadoop streaming to create a job that outputs the most common word that starts with a vowel and the most common word that starts with a consonant. The output should also include the number of times the words appear. In addition, use counters to output the total number of unique words that start with vowels and that start with consonants.

   One way to solve this problem would be to use vowel/consonant as the key and pack the words and their counts into the values. Doing that will make using a combiner harder (should we choose to do that later), so the values emitted from the map phase should contain only numeric word counts. *Hint:* use the field separator settings (http://hadoop.apache.org/docs/r0.18.1/streaming.html#Customizing+the+Way+to+Split+Lines+into+Key%2FValue+Pairs).

What to hand-in: Upload the source code to http://snap.stanford.edu/submit-cs246h/. You may upload as many time as you like.