

Problem Set 1

Due 12:30pm January 29, 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. Take the original WordCount job and add a combiner to it. What's different about the output? What's different about the counters that are printed after the job completes?
2. Write a MapReduce 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. 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:* try implementing a custom partitioner. *Python hint:* if you're doing this with Python, you should look into 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.