CS246: Mining Massive Datasets Hadoop Lab **Winter 2014**

Problem Set 1

Due 12:30pm January 22, 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. Write a Hadoop MapReduce program which outputs the full dataset in all caps. In your driver code, set the number of reducers to 2. Note that the results should not include anything other than the original input text in all caps. Download the output using the command:

   ```
   hadoop fs -getmerge output
   ```

   where `output` is the output directory created by your job. Compare the results with the original. What happened and why?

2. Write a MapReduce program to output **only** the single most common bigram (pair of adjacent words) in the dataset. Split words on whitespace.

3. Write a MapReduce program to output only the lines that contain the word "torture". What role does the reducer play in your job?

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