**SI 618 Fall 2017 Lab 2 - mrjob (20 points)**

This lab is to familiarize you with the process of writing MapReduce code and running it locally on your laptop.

In the lab, you need to compute the bigram counts. A bigram is every sequence of two adjacent words in a string of words. For the sake of simplicity, we will treat each line in the input text file as a string. For example, if the text is:

*one two three four three four five six*

Then we should have these unique bigrams and counts:

"*one two*" 1 "*two three*" 1 "*three four*" 2 "*four five*" 1 "*five six*" 1

To get started, download the 'si618\_f17\_lab2.py’ file, and rename si618\_f17\_lab2.py' as si618\_f17\_lab2 \_youruniquename.py'.

**Part 0. Finding the input data and installing mrjob**

The input data for lab 2 is in on the course Canvas site. Find and download the file books.zip, and unizip it. This contains the text of several books from the Gutenberg Project. There should be 10 items in the directory.

If you don’t already have the mrjob Python 3 module installed, you should install it by running

$ pip install mrjob

or

$ pip3 install mrjob

**Part 1. MapReduce**

Your regular expression should match just like shaded areas below.

Project Gutenberg's Frankenstein, by Mary Wollstonecraft (Godwin) Shelley

re-use this eBook. or online at www.gutenberg.net

Author: Mary Wollstone\_craft (Godwin) Shelley

Release Date: June 17, 2008 [EBook #84]

Add code to si618\_f17\_lab2\_youruniquename.py where specified. Then, run this script locally (on your computer) on the input files in the books directory.

To run the code, you should enter the following:

$ python si618\_f17\_lab2\_youruniquename.py ./books -o si618\_f17\_lab2\_output

or

$ python3 si618\_f17\_lab2\_youruniquename.py ./books -o si618\_f17\_lab2\_output

If your script runs successfully, the script will create a directory called si618\_f17\_lab2\_output containing files with names like ‘part-000000’, ‘part-00001’, etc. (the exact number of files depends on the number of cores used to run the mrjob code). Now, to concatenate those into a single file, run

$ cat si618\_f17\_lab2\_output/part\* > si618\_f17\_lab2\_output\_youruniquename.txt

Note that the order of lines in your bigram output file may be different from the desired output, and that is OK.

As an additional note, you could produce this text output in a single line when running your script by adding > si618\_f17\_lab2\_output\_youruniquename.txt to the command used to run the script above:

python si618\_f17\_lab2\_youruniquename.py ./books -o si618\_f17\_lab2\_output > si618\_f17\_lab2\_output\_youruniquename.txt

This would redirect the output from the terminal to the text file specified after the ‘>’.

**What to submit:**

Submit a zip file named si618\_f17\_lab2\_youruniquename.zip containing your Python source code file si618\_f17\_lab2\_youruniquename.py and your merged output file si618\_f17\_lab2\_output\_youruniquename.txt.