**SI 618 Fall 2019 Homework 5 (100 points)**

Data to be used in this homework: On the Hadoop cluster, I have put the following file in HDFS:

hdfs:///var/umsi618/hw5/review.json   
hdfs:///var/umsi618/hw5/business.json

These files were downloaded from <http://www.yelp.com/dataset_challenge> (you cannot share the content with others without going through the approval procedure yourself). The format of the data is explained in the “Notes on the Dataset” section at <http://www.yelp.com/dataset_challenge>.

*Note: You do not need to download the Yelp dataset yourself as it is already put into HDFS on the Hadoop cluster.*

**Traveling Yelp Users**

The goal of this question is to find out the distribution of the number of distinct cities that Yelp users wrote reviews in. We can imagine that some Yelp users travel a lot so they wrote reviews for businesses in a bunch of cities, but most Yelp users probably only wrote reviews for businesses in one single city. Is this true?

To answer this question, you are going to use Spark to join these two data sets together and produce a breakdown of the Yelp users by the number of distinct cities they wrote **greater than one star** reviews in.

You code should save the result in an CSV file with two columns: “cities” and “yelp users”

For example:

cities,yelp users

1,1012388  
2,221758

This means that 1012388 yelp users wrote reviews for businesses in once city only.

Your results should be exactly the same as the provided si618\_hw5\_desired\_output.csv.   
Save your file as si618\_hw5 \_youruniquename.csv.

Now that you have figured out the overall pattern, see if the pattern is different for positive and negative reviews. To do this, I want you to restrict your analysis to reviews **with more than 3 stars** *(good reviews)* in one case and reviews with **less than three stars** *(bad reviews)* in the other case. Your results should look like:

* si618\_hw5\_desired\_output\_goodreview.csv
* si618\_hw5\_desired\_output\_badreview.csv

Save your files as si618\_hw5 \_youruniquename\_goodreview.csv and si618\_hw5 \_youruniquename\_badreview.csv respectively.

Your Spark code should run as a standalone application on the Cavium cluster.

Hint: You can use the histogram() function to calculate the breakdown. See: <https://spark.apache.org/docs/1.5.0/api/python/pyspark.html#pyspark.RDD.histogram>

**What to submit:**

* Submit a **zip file** named si618\_hw5\_youruniqname.zip containing:
  + si618\_hw5 \_youruniquename.py
  + si618\_hw5\_youruniquename.csv
  + si618\_hw5 \_youruniquename\_goodreview.csv
  + si618\_hw5 \_youruniquename\_badreview.csv