**SI 618 Fall 2017 Homework 4 (100 points)**

Data to be used in this homework: On the Fladoop cluster, I have put the following two files in HDFS:

hdfs:///var/si618f17/yelp\_academic\_dataset\_business\_updated.json  
hdfs:///var/si618f17/yelp\_academic\_dataset\_review\_updated.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 that you do not need to download the Yelp dataset yourself as it is already put into HDFS on the Fladoop cluster.

The goal of this homework 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 SparkSQL to join these two data sets together and produce a breakdown of the Yelp users by the number of distinct cities they wrote reviews in.

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

cities,yelp users

1,280598

means that 280598 yelp users wrote reviews for businesses in once city only.

Your results should be exactly the same as the provided hw4\_desired\_output\_allreview.csv. Save your file as si618\_f17\_hw4\_output\_allreview\_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 in one case and reviews with less than three stars in the other case. Your results should look like hw4\_desired\_output\_goodreview.csv and hw4\_desired\_output\_badreview.csv for these two cases respectively. Save your files as si618\_f17\_hw4\_output\_goodreview\_ youruniquename.csv and si618\_f17\_hw4\_output\_badreview\_youruniquename.csv respectively.

Your Spark code should (1) use SparkSQL, (2) run as a standalone application on the Fladoop cluster, i.e., it should run by issuing this command on Fladoop login node:

spark-submit --master yarn-client --queue si618f17 si618\_f17\_hw4\_uniqname.py

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>

**What to submit:**

Submit a zip file named si618\_f17\_hw4\_uniqname.zip containing your Python source code file and 3 csv files.