SI 601 Winter 2016 Homework 6 (100 points)

Due at 5:30pm on Wednesday, Feb. 24, 2016

On the Fladoop cluster, I have put the following two files in HDFS:

hdfs:///user/yuhangw/yelp_academic_dataset_business.json
hdfs:///user/yuhangw/yelp_academic_dataset_review.json

These files were downloaded from http://www.yelp.com/dataset_challenge

The format of the data is explained in the "Notes on the Dataset" section at http://www.yelp.com/dataset_challenge

Chaplin

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 assignment 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 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 si601f15hw6_desired_output.csv.

Your Spark code should 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 si601w16 si601_w16_hw6_ youruniqname.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 si601_w16_hw6_youruniqname.zip containing your Python source code file and the your output CSV file.