SI 601 Winter 2016 Homework 5 (100 points)

Due at 5:30pm on Wednesday, Feb. 10, 2016

In this homework assignment, you will use PySpark to analyze the Yelp's Academic Dataset (http://www.yelp.com/academic_dataset). You do not need to download the dataset because I already put it in HDFS on the Fladoop cluster.

```
[yuhangw@flux-login2 ~]$ hadoop fs -ls /user/yuhangw/yelp_academic_dataset.json
-rw-r--r--   3 yuhangw hadoop  346446934 2016-02-01 10:52 /user/yuhangw/yelp_academic_dataset.json
```

The format of the data is explained at http://www.yelp.com/academic_dataset. For this assignment, we only care about the Business Objects in the data.

The goal is to compute the number of businesses, total review count, and average star rating for each neighborhood in each city. If a business has multiple neighborhoods, its review count and stars should be attributed to all of the neighborhoods. If the neighborhoods list is empty, then we will use 'Unknown' as the name of the neighborhood.

Your final result should be a TSV file that is the same as the provided si601w16hw5_output_desired_output.tsv file.

In this desired output file, each row contains 5 columns, which are separated by a tab. For example, this row

Ann Arbor	Downtown Ann Arbor	273	7137	3.66117216117

means the neighborhood of “Downtown Ann Arbor” in the city of “Ann Arbor” has 273 businesses, and their total review count is 7137, and their average star rating is 3.66.

The rows in the output file should be sorted in alphabetical order of the city names, and the neighborhoods in each city are sorted by the number of businesses in decreasing order.

You MUST use Spark to do this homework. A non-Spark solution will not get any credit.

HINT: You can modify from the provided example code spark_avg_stars_per_category.py

What to submit

Submit a zip file named si601_w16_hw5_youruniquename.zip containing your PySpark source code and your output TSV file.