**SI 618 Fall 2020 Homework 5 (100 points)**

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

hdfs:///data/umsi618f20/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 that you do not need to download the Yelp dataset yourself as it is already put into HDFS on the Hadoop cluster.*

**Highly Rated City Businesses**

Here you will use hdfs:///data/umsi618f20/hw5/business.json

The goal is to compute the number of businesses, total review count, and number of 3-star or higher reviews for each category in each city. If a business has multiple categories, its review count and stars should be attributed to all of the categories. If the category list is empty, then we will use 'Unknown' as the name of the category.

Your final result should be a TSV file that is the same as the provided   
si618\_hw5 \_desired\_output.tsv file.

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

Airdrie Auto Repair 13 49 11

This means the category of “Auto Repair” in the city of “Airdrie” has 13 businesses, and their total review count is 49, and 11 of the reviews are 3 stars or more.

The rows in the output file should be sorted in alphabetical order of the city names, and the categories in each city are sorted by the number of total review counts that are 3 stars or more in decreasing order. Note that there are some cities with names that may be data entry errors (such as “110 Las Vegas”); to match the desired output you should leave these cities in your output and not perform any filtering to remove cities with strange/incorrect names or letter casing.

**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. Save it as “si618\_hw5 \_youruniquename.py”

**What to submit:**

Submit a zip file named si618\_hw5\_youruniqname.zip containing:

* Your Python source code: si618\_hw5\_youruniquename.py
* The tsv outpout file: si618\_hw5 \_output\_youruniquename.tsv