**SI 618 Fall 2020 Homework 6 (100 points)**

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

hdfs:///data/umsi618f20/hw6/review.json   
hdfs:///data/umsi618f20/hw6/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.*

**Distinguishing Yelp Users Rating Behaviors**

Reviewers can have different thresholds for defining something as good. Some reviewers may give higher ratings across all businesses while others may give lower ones. We can identify the average rating businesses get by accounting for this behavior. The goal of this question is to understand preferences of reviewers and normalize this when rating a business.

**(20 points)** To answer this question, you are going to use Spark SQL to join these two data sets together and produce a breakdown of normalized rating by taking user star rating (stars from reviews.json) and subtracting it by the average business star rating (stars from business.json).

Ex:

SELECT b.business\_id, (r.stars-b.stars)

We want to get all business ids that were reviewed by users and their respective normalized rating.

Your code should be a two column tsv file that looks something like this:

ujmEBvifdJM6h6RLv4wQIg, -1.5  
NZnhc2sEQy3RmzKTZnqtwQ, -0.5

This means that for business *ujmEBvifdJM6h6RLv4wQIg* a specific yelp reviewer has a normalized rating of -1.5 for that business.

Your results should be exactly the same as the provided **hw6\_desired\_output\_1.tsv**.   
Save your file as **si618\_hw6\_output\_1\_youruniquename.tsv**.

**(40 points)** Now that you have the normalized rating for every business that yelp users rated, can we find the average normalized rating for each business? For example, can we find the average normalized rating for business ujmEBvifdJM6h6RLv4wQIg? Please return your findings in ascending order of the average normalized rating. Your output should match:

* **hw6\_desired\_output\_2.tsv**

Save your files as **si618\_hw6\_output\_2 \_youruniquename.tsv**

**(40 points)** Lastly, can we find the average normalized ratings on a city level? In other words, what is the average normalized rating of all businesses in a city like Chicago? Please return that in descending order of average normalized rating for each city. Your output should match:

* **hw6\_desired\_output\_3.tsv**

Save your files as **si618\_hw6\_output\_3 \_youruniquename.tsv**

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

**What to submit:**

* Submit a **zip file** named si618\_hw6\_youruniqname.zip containing:
  + si618\_hw6 \_youruniquename.py
  + si618\_hw6\_output\_1\_youruniquename.tsv
  + si618\_hw6\_output\_2\_youruniquename.tsv
  + si618\_hw6\_output\_3\_youruniquename.tsv