SI 618 Fall 2021 Homework 6 (100 points)

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

hdfs:///var/umsi618f21/hw6/yelp_academic_dataset_business.json
hdfs:///var/umsi618f21/hw6/yelp_academic_dataset_review.json

This file was downloaded from the Kaggle Yelp dataset.

Note that 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 of 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.

(40 points) First, find normalized ratings of individual users by using the Z-score formula as follows. To simplify the analysis, consider the normalized rating to be 0 when the denominator of the formula is 0.

$$Normalized\ Rating = (Rating - Average\ Rating)/(Std.Deviation\ in\ Rating)$$

Next, generate the average normalized rating for each business by using the normalized ratings of users. Sort the businesses by descending order of normalized rating and return only the top 100 businesses.

The output should have the format of **business_id<tab>normalized_rating**

Your results should be exactly the same as the provided **hw6_desired_output_1.tsv**. Save your file as **unique_name_si618_hw6_output_1.tsv**.

(30 points) Now that you have the average normalized rating for every business (not just the 100 you output for part 1), find the average normalized rating for businesses by city. For example, what is the average normalized rating of businesses in Chicago?' Note that, as with homework 5, you **should not** do any filtering or reformatting of the city name to fix any issues.

Sort the cities in the descending order of their average normalized business rating. The output should have the following format: **city<tab>average_business_rating**.

The output should be the same as the provided **hw6_desired_output_2.tsv** Save your files as uniquename**_si618_hw6_output_2.tsv**

(30 points) Finally, consider that the community of users find some reviews more useful than others. Recalculate the city level average normalized business rating by considering user reviews that at least one other person from the community has found useful (hint: the "useful" field in each review gives the number of people who found a review useful). Note that you don't need to recalculate the normalized ratings for each user using useful ratings, only, when calculating average ratings for cities, limit already calculated user normalized ratings to useful ones.

Sort the cities in the descending order of their average normalized business rating. The output should have the following format: **city<tab>average_business_rating**.

The output should be the same as the provided **hw6_desired_output_3.tsv** Save your files as uniquename**_si618_hw6_output_3.tsv**

Your Spark code should run as a standalone application on the Cavium cluster. You MUST use **SparkSQL** to do this lab. Other solutions will not get any credit.

What to submit:

- si618_hw6 _youruniquename.py
- si618_hw6_output_1_youruniquename.tsv
- si618_hw6_output_2_youruniquename.tsv
- si618_hw6_output_3_youruniquename.tsv

Submit these as separate files and not as a zipped folder. It helps the IAs to grade your work more efficiently