CS 4371.501 Introduction to Big Data Management and Analytics

Homework 4

In this homework, you are asked to use Spark (any Spark library, including Spark SQL) for solving the following questions about **Yelp Dataset**. Before presenting the questions, a brief description of the dataset is given below:

business.csv file contains basic information about local businesses, and it contains the following columns:

- *'business_id'*: (a unique identifier for the business)
- *'full_address'*: (localized address),
- *'categories'*: [(localized category names)]

review.csv file contains the star rating given by a user to a business. Use user_id to associate this review with others by the same user. Use business_id to associate this review with others of the same business. This file contains the following columns:

- *'review_id'*: (a unique identifier for the review)
- *'user_id'*: (the identifier of the reviewed business),
- *'business_id'*: (the identifier of the authoring user),
- *'stars'*: (star rating, integer 1-5), the rating given by the user to a business.

user.csv file contains aggregate information about a single user. This file contains the following columns:

> *'user_id'*: (unique user identifier),
>
> *'name'*: (first name, last initial, like 'Matt J.'), this column has been made anonymous to preserve privacy,
>
> *'url'*: url of the user on Yelp.

Notes:

1. **Delimiter "::" is column separator for all the files shown above.**

2. **All the sample outputs shown in the questions below are merely used for illustration. In fact, they were computed over the Yelp dataset.**

(Question 1) Compute the total number of posted reviews and average number of stars for each business_id. Display these aggregated figures.

Sample output:

```
business_id          NumOfReviews              AvgRating
xdf1234444...        517                       4.34
………………………………………………………………………………..
```

(Question 2) Compute the total number of posted reviews and average number of stars for each State. Display these aggregated figures sorted by the average ratings.

Sample output:

```
State      NumOfReviews              AvgRating
NY         9222                      4.53
CA         10425                     4.40
………………………………………………………………………………
```

(Question 3) List the *'user id'*, *'name'* and average *'rating'* of users that reviewed businesses classified as “*Colleges & Universities*” in list of categories. Note that the average rating must be computed over all the ratings posted by the user ids and not only the ratings related to businesses classified as “*Colleges & Universities*”.

Sample output:

UserID	**Name**	**AvgRating**
Tpmvufw1eea1DrjLAY2jLg	Theodore J.	4.0

………………………………………………………………

(Question 4) List the business_id, full address and categories of the Top 10 businesses located in “NY” using the average ratings. Display the results sorted by average rating.

Sample output:

BusinessID	**FullAddress**	**Categories**	**AvgRating**
xdf1234444....	CA 91711…	List['Local Services', 'Carpet Cleaning']	5.0

………………………………………………………………………………

(Question 5) List the 'user id', 'name' and average 'rating' of users that reviewed businesses located in more than one State.

UserID	**Name**	**AvgRating**
Tpmvufw1eea1DrjLAY2jLg	Theodore J.	4.0

………………………………………………………………

Submission:

You have to upload your submission via e-learning before the due date.

Please upload the following to eLearning:

1. source files (notebook, .py or .scala)
2. output of your program

Note that, for this homework, you are also allowed to use Databricks.