In this project, you will use spark (pyspark or scala) to solve the following problems.

**Q1.**

In this question, you will learn how to solve problems using **Apache Spark**. Please use Apache Spark to derive some statistics from **Yelp Dataset.**

**Data set info:**

The dataset files are as follows and columns are separate using ‘**::**’

**business.csv.**

**review.csv.**

**user.csv.**

**Data set Description.**

The data set comprises of **three** csvfiles, namely user.csv, business.csv and review.csv.

**Business.csv** file contain basic information about local businesses.

**Business.csv** file contains the following columns "business\_id"::"full\_address"::"categories"

'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.

**review.csv** file contains the following columns "review\_id"::"user\_id"::"business\_id"::"stars"

'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 across all of Yelp

**user.csv file** contains the following columns "user\_id"::"name"::"url"

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

**NB: :: is Column separator in the files.**

**List the 'user id' and 'rating' of users that reviewed businesses located in “Stanford”**

Required files are 'business' and 'review'.

**Sample output**

|  |  |
| --- | --- |
| **User id** | **Rating** |
| 0WaCdhr3aXb0G0niwTMGTg | 4.0 |

**Q2:**

**List the business\_id , full address and categories of the Top 10 businesses using the average ratings.**

This will require you to use **review.csv** and **business.csv files.**

**Sample output:**

**business id full address categories avg rating**

xdf12344444444, CA 91711 List['Local Services', 'Carpet Cleaning'] 5.0