CS 6350- ASSIGNMENT 2

Please read the instructions below before starting the assignment.

- This assignment consists of 5 questions. You should create an Eclipse project with a different class for each of the questions.

- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6350/CS6350_CoverPage.docx

- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page. Only one submission per team is required.

- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.

- Please ask all questions on Piazza, and not through email to the instructor or TA.

Assignment 2

In this lab, you will learn how to solve problems using Map Reduce. You will use Hadoop map-reduce to derive some statistics for the **Yelp Dataset.**

The dataset files are located in HDFS in the following path,

/yelp/business/business.csv
/yelp/review/review.csv
/yelp/user/user.csv

If somehow the files disappear from the above HDFS location, you can also download them from:

http://www.utdallas.edu/~axn112530/cs6350/yelp/

What to submit:
- You should create an Eclipse project with different classes for each of the questions
- Your project should build successfully and generate a jar file
- Your project should run on the UTD cluster
- You should copy the output for each of the questions and paste them in a text file which should be submitted.
- Create a zip file for the project and the output and submit on eLearning

---

Dataset Description.
The dataset comprises of **three** csv files, namely user.csv, business.csv and review.csv. Note that some of the content, such as id fields are encoded. Note that the files are separated by "^" character.

1. 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)]

2. 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

3. 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

Q1. For all the businesses that are located in "Palo Alto", output their full address and also how many businesses are in each address. You can use the full_address column as the filter column.
(An example of how to do this is in the file CountYelpBusiness.java).

Q2. Modify Q1 to output business id and full_address of Restaurants that are located in the state of NY.

Q3. You would like to find the top 10 zip codes where the most businesses are located. To accomplish this, you will emit the following (K,V) pair from mapper (ZipCode, 1). Then in the reducer, you will sort by the value and emit the top 10 elements.
Example code for the topN values for the wordcount problem is given in the class TopN. You can get some hints from that class.

Q4. Find the top ten rated businesses using the average ratings. Recall that star column in review.csv file represents the rating.

Please answer the question by calculating the average ratings given to each business using the review.csv file. You can reuse part of the logic for sorting by values from Q3.

Sample Output:

eebUeWSJDlmtz80tT2kDuA	5.0
H7VLT9-UbaDVKbxfLAMqwg	5.0
dLJgjRFphvHoQQsC9tEyTQ 5.0	5.0

Q5. Modify Q4 to find out the 10 businesses that have received the lowest average ratings.
*** Hint: You just have to output hashmap in reverse order and stop at counter value of 10 ***