CS6350
Big data Management Analytics and Management
Summer 2015
Homework 1
Submission Deadline: 26th June, 2015

In this homework, you will learn how to solve problems using Map Reduce. Please apply

Hadoop map-reduce to derive some statistics from **Yelp Dataset(**Original source https://www.yelp.com/academic_dataset**)**. You can find the dataset in elearning. Please copy the data into your hadoop cluster and use it as input data.

You can use the **put** or **copyFromLocal** HDFS shell command to copy those files into your HDFS directory.

In class there will be brief demo/ discussion about that.

Dataset Description.
The dataset comprises of a **single** csv file, **data.csv** that contains 3 types of entities, namely users, businesses and reviews. Records for each entity are distinguished by the '**type**' column.

The "type" column determines the type of an entity a row represents. For example,

if **type is business**, then that **row contains business data,**

if **type is user** , then the **row contains user data, and**

likewise if **type is review, then the row contains review data.**

The csv file has 24 columns, namely

Column id : Name of Column
Column 0 :review_id
Column 1: text
Column 2: business_id
Column 3: full_address
Column 4: schools
Column 5: longitude
Column 6: average_stars:: //this is for the business entity type only
Column 7: date
Column 8: user_id
Column 9: open
Column10: categories
Column11: photo_url

Column12: city
Column13: review_count
Column14: name
Column15: neighborhoods
Column 16: url
Column 17: votes.cool
Column 18: votes.funny
Column 19: state
Column 20: stars:: //this is for review entity type only
Column 21: latitude
Column 22: type
Column 23: votes.useful

The columns specific to each entity type is shown below:

Business Entities

Business objects contain basic information about local businesses.
{
 'type': 'business',
 'business_id': (a unique identifier for this business),
 'name': (the full business name),
 'neighborhoods': (a list of neighborhood names, might be empty),
 'full_address': (localized address),
 'city': (city),
 'state': (state),
 'latitude': (latitude),
 'longitude': (longitude),
 'stars': (star rating, rounded to half-stars),
 'review_count': (review count),
 'photo_url': (photo url),
 'categories': [(localized category names)]
 'open': (is the business still open for business?),
 'schools': (nearby universities),
 'url': (yelp url)
}

Review Entities

Review objects contain the review text, the star rating, and information on votes Yelp users have cast on the review. 'user_id' will be used to identify the users who provide the review . Similarly 'business_id' will be used to associate a review with a particular business entity.

{
'type': 'review',
'business_id': (the identifier of the reviewed business),
'user_id': (the identifier of the authoring user),
'stars': (star rating, integer 1-5),
'text': (review text),
'date': (date, formatted like '2011-04-19'),
'votes.useful': (count of useful votes),
'votes.funny': (count of funny votes),
'votes.cool': (count of cool votes)
}
}

User Entities

User objects contain aggregate information about a single user across all of Yelp

{
'type': 'user',
'user_id': (unique user identifier),
'name': (first name, last initial, like 'Matt J.'),
'review_count': (review count),
'average_stars': (floating point average, like 4.31),
'votes.useful': (count of useful votes across all reviews),
'votes.funny': (count of funny votes across all reviews),
'votes.cool': (count of cool votes across all reviews)
}
}

After being familiar with the data - you are required to **write efficient Hadoop Map-**

Reduce programs in Java to find the following information ::

Q1:
Q 1 a: Count the total number of reviews,
Q 1 b: Count total number of users
Q 1 c: Count total number of business entities in the data.csv file.

This demonstrates the use of MapReduce to filter and count data.

Sample output

review 245
user 200
business 347

Q2.
List each business Id that are located in “Palo Alto” using the full_address column as the filter column.

This also demonstrates the use of Hadoop to filter data.

Sample output:

23244444
232ewe33

Q3

Find the top ten rated businesses using the average ratings.
The star column represents the rating.

Please answer the question by calculating the average ratings given to each business using the review entity rows. Do not use the already calculated ratings (average_stars) contained in the business entity rows.

This will require you to use entity of “type” review**.**

Sample output:
business id
xdf12344444444

Q4:
Please use reduce side join and job chaining technique to answer question 4.

List the business_id , full address and categories of the Top 10 businesses using the average ratings.

This will require you to use entity of “type” **review** and **business.**

Important:
Please note that some business ids do not have full entry in the business type rows. Please list the top 10 businesses that have entries in the business type rows.

Sample output:

business id	full address	categories	avg rating
xdf12344444444,	CA 91711	['Local Services', 'Carpet Cleaning']	5.0

Q5 Please use Map side join technique to answer this question

Load all business rows into the distributed cache. **There are only 78 rows that contains business entity type.**

List the 'user id' and 'review text' of users that reviewed businesses located in Stanford

Required entity type is 'business' and 'review'.

Sample output

User id	Review Text
0WaCdhr3aXb0G0niwTMGTg	We hired Stanford's Bartender for a private movie screening party and will definitely use them again for all our events in the future.

Submission ::
You have to upload your submission via e-learning before due date.
Please upload the following to eLearning:
1. The jar files, one for each problem.
2. Java files which have the source code.
3. An output of your program
4. ***A Readme text file about how to run your jar file. Give the command to run your jar file.