CS6350
Big data Management Analytics and Management
Fall 2017
Homework 1
Submission Deadline: 6th October, 2017

In this homework, you will use hadoop/mapreduce to solve the following problems.

Q1

Write a MapReduce program in Hadoop that implements a simple "Mutual/Common friend list of two friends". The key idea is that if two people are friend then they have a lot of mutual/common friends. This question will give any two Users as input, output the list of the user id of their mutual friends.

For example,
Alice's friends are Bob, Sam, Sara, Nancy
Bob's friends are Alice, Sam, Clara, Nancy
Sara's friends are Alice, Sam, Clara, Nancy

As Alice and Bob are friend and so, their mutual friend list is [Sam, Nancy]
As Sara and Bob are not friend and so, their mutual friend list is empty. (**In this case you may exclude them from your output**).

Input:
Input files
1. soc-LiveJournal1Adj.txt located in /socNetData/networkdata in hdfs on cs6360 cluster (I will also provide this file so that you can test it locally)

The input contains the adjacency list and has multiple lines in the following format:
<User><TAB><Friends>

Here, <User> is a unique integer ID corresponding to a unique user and <Friends> is a comma-separated list of unique IDs (<User> ID) corresponding to the friends of the user. Note that the friendships are mutual (i.e., edges are undirected): if A is friend with B then B is also friend with A. The data provided is consistent with that rule as there is an explicit entry for each side of each edge. So when you make the pair, always consider (A, B) or (B, A) for user A and B but not both.

Output: The output should contain one line per user in the following format:
<User_A>, <User_B><TAB><Mutual/Common Friend List>

where <User_A> & <User_B> are unique IDs corresponding to a user A and B (A and B are friend). < Mutual/Common Friend List > is a comma-separated list of unique IDs corresponding to mutual friend list of User A and B.

Please find the above output for the following pairs.

(0,4), (20, 22939), (1, 29826), (6222, 19272), (28041, 28056)

Q2.
Please answer this question by using dataset from Q1.
Find friend pairs whose common friend number are within the top-10 in all the pairs. Please output them in decreasing order.
Output Format:
<User_A>, <User_B><TAB><Mutual/Common Friend Number>

Q3.

In this question, you will apply Hadoop map-reduce to derive some statistics from **Yelp Dataset.**

---------- **Data set Info** ----------
The dataset files are as follows and columns are separated using '**::**'
business.csv.
review.csv.
user.csv.

Dataset Description.
The dataset comprises of **three** csv files, 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 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.**

Please use reduce side join and job chaining technique to answer this problem.

Sample output:

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

Q4.
Use **Yelp Dataset**

List the 'user id' and 'rating' of users that reviewed businesses located in "Palo Alto"
Required files are 'business' and 'review'.

Please use In Memory Join technique to answer this problem.
Hint: Please load all data in business.csv file into the distributed cache.

Sample output

User id	**Rating**
0WaCdhr3aXb0G0niwTMGTg	4.0

Submission Instructions:

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.