CS6350
Big data Management Analytics and Management
Spring 2023
Homework 1
Submission Deadline: March 3rd, 2023, 11:59pm

In this homework, you will be using Hadoop/MapReduce to analyze social network data.

Q1: Write a MapReduce program in Hadoop that finds **Mutual/Common friend list of two friends specifically for the pairs (0,1), (20, 28193), (1, 29826), (6222, 19272), (28041, 28056)**. The key idea is that if two people are friend then they have a lot of mutual/common friends. This program will find the common/mutual friend list for them.

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

Input:

1. mutual.txt

The input contains the adjacency list and has multiple lines in the following format:
<User><TAB><Friends>
Hence, each line represents a particular user’s friend list separated by comma.

2. userdata.txt

The userdata.txt contains dummy data which consist of
column1 : userid
column2 : firstname
column3 : lastname
column4 : address
column5: city
column6 :state
column7 : zipcode
column8 :country
column9 :username
column10 : date of birth.

Here, <User> is a unique integer ID corresponding to a unique user and <Friends> is a comma-separated list of unique IDs corresponding to the friends of the user with the unique ID <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 pair 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.

Q2: Please use **in-memory join at the Mapper** to answer the following question.
For each user print User ID and average age till 01/01/2023 of direct friends of this user. Note that the userdata.txt will be used to get the extra user information and cached/replicated at each mapper.

Output format:
<User><TAB><Average age of direct friends>

Sample output:

```
1234    60
```

Q3: Please **use in-memory join at the Reducer** to answer the following question.
Given any two Users (they are friends) as input, output a list containing the dates of birth of all their mutual friends and the age of the youngest friend in days till 01/01/2023.

Output format:
<User_A>, <User_B><TAB>< List of DOBs [date$_1$, date$_2$, ... date$_n$], Age of youngest friend in days until 01/01/2023>

Sample output:

```
1234    4312    [10/12/1994, 01/03/1996, 11/11/1995], 9860
```