Assignment 1: Big Data Processing

Name: Waghe Shubham Yatindra
Roll No.: **13MF3IM17**

Algorithm for M/R program

1. Take in all inputs and process them in required format
2. For each line -> emit tuple as follows:
   (<id>, (<user>, <identifier>)) where **id**: is key, **user** is related to **id** by **identifier**
   Identifier: 1 for one mutual friend and **-1** if already friend. Eg: id f1,f2,....fn
   a) (id, (f1, -1) [already friends for all such f]
   b) (f1, (f2, 1)) and (f2, (f1,1)) [for all such **f1** and **f2** having mutual friend as **id**]
3. Reducer reduces counts of mutual friends by adding for all friends, **-1** if already friends
4. Sorted results according to mutual friend count and **id** (in case of tie, ascending) is given as output.

Submissions

1. Hadoop Java program - Logic uses **id** of mutual friend and then counts the numbers and **-1** if already friends.
2. Python Apache Spark program using pyspark - Logic uses **1** for each mutual friend and **-1** for already friends.

Results

924 439,2409,6995,11860,15416,43748,45881
8941 8943,8944,8940
8942 8939,8940,8943,8944
9019 9022,317,9023
9020 9021,9016,9017,9022,317,9023
9021 9020,9016,9017,9022,317,9023
9022 9019,9020,9021,317,9016,9017,9023
9990 13134,13478,13877,34299,34485,34642,37941
9992 9987,9989,35667,9991
9993 9991,13134,13478,13877,34299,34485,34642,37941