Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 3

Due: Feb 8, 2024(100 points)

Yu Wang
Git User Name: titojojo
Email: wang.yu25@northeastern.edu
https://github.com/Titojojo/CS6220-Data-Mining

People You Might Know

This Spark program analyses social network data and recommends the friends a user might know. There are several main steps in the program:

1. Parse input data: Read and parse the input .txt file to a list of [user, friends] pairs.
2. Create friend pairs: From each user's network, generate friend pairs. Mark already connected friends to negative infinity.
3. Filter results and get potential friend pairs: Use 'reduceByKey' to aggregate the count of mutual friends. Filter out existing friends.
4. Sort results and output the top N.

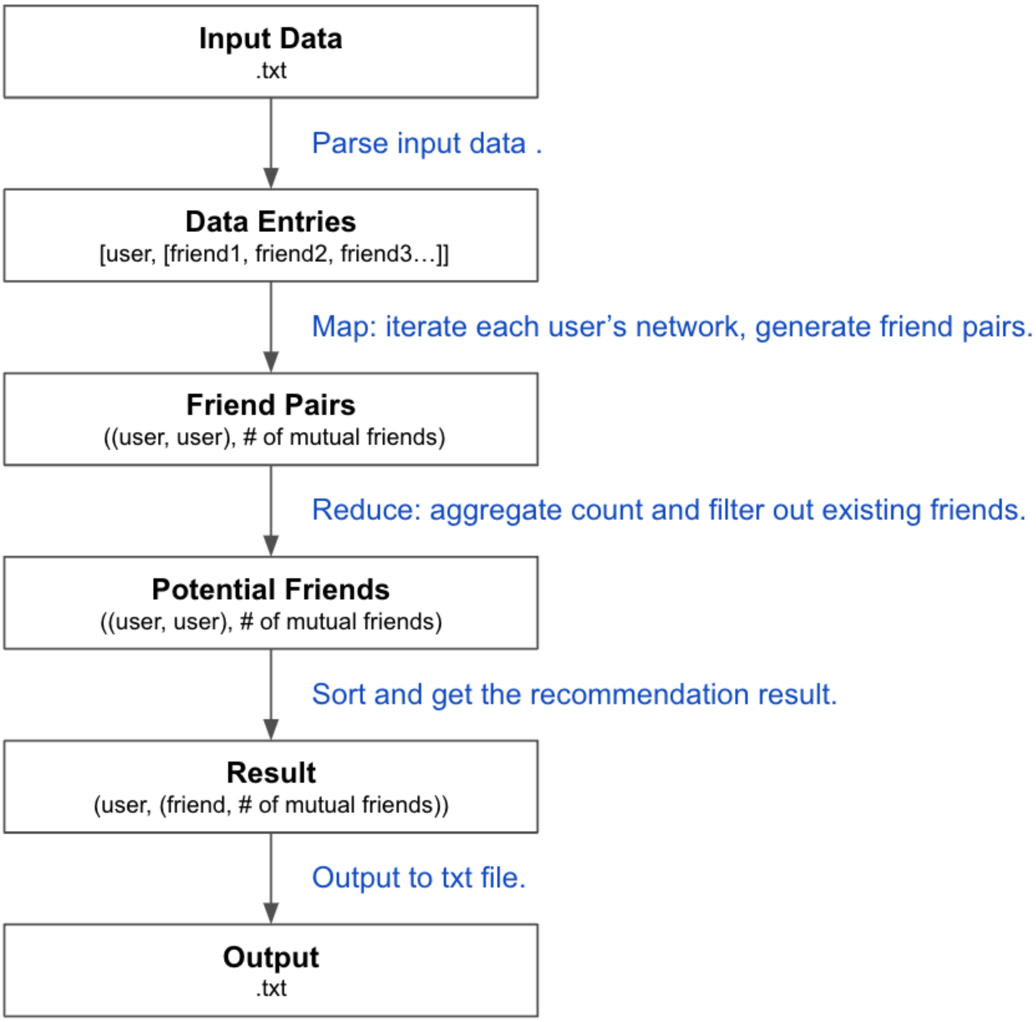
Input Data
.txt
Parse input data .
Data Entries
[user, [friend1, friend2, friend3…]]
Map: iterate each user's network, generate friend pairs.
Friend Pairs
((user, user), # of mutual friends)
Reduce: aggregate count and filter out existing friends.
Potential Friends
((user, user), # of mutual friends)
Sort and get the recommendation result.
Result
(user, (friend, # of mutual friends))
Output to txt file.
Output
.txt

Figure 0.1: Pipeline Sketch

Recommendation Results

924: 11860,15416,2409,43748,439,45881,6995

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