Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 3

Due: February 9, 2024(100 points)

Dharun Suryaa Nagarajan
https://github.com/dharun4772/CS6220/tree/main/Assignment%203
nagarajan.dh@northeastern.edu

Map Reduce in Spark

Write a Spark program that implements a simple “People You Might Know” social network friendship recommendation algorithm. The key idea is that if two people have a lot of mutual friends, then the system should recommend that they connect with each other

Code and approach explanation:

The given code is a PySpark implementation for finding mutual friends in a social network dataset and generating recommendations based on those mutual friendships.
It starts by loading data from a file containing user-friend relationships. The code then processes the data to create pairs of friends from the friends array list, filters out the redundant pairs from the friend pairs formed using the greater than logic and identifies mutual friends that commonly occur.
After grouping the results, it constructs a DataFrame and generates recommendations by sorting and selecting at least 10 mutual friends if present for each user.
Finally, the results are written to an output file, and examples for specific users (924, 8941, 8942, etc.) are printed.
The code efficiently utilizes PySpark's distributed computing capabilities to handle large-scale social network datasets.

Recommendations for the users with following user IDs:

1. 924 : 2409, 6995, 11860, 15416, 43748, 45881
2. 8941 : 8943, 8944
3. 8942 : 8943, 8944
4. 9019 : 9022, 9023
5. 9020 : 9021, 9022, 9023
6. 9021 : 9022, 9023
7. 9022 : 9023
8. 9990 : 13134, 13478, 13877, 34299, 34485, 34642, 37941
9. 9992 : 35667
10. 9993 : 13134, 13478, 13877, 34299, 34485, 34642, 37941

Diagram of pipeline description: