Northeastern University, Khoury College of Computer Science

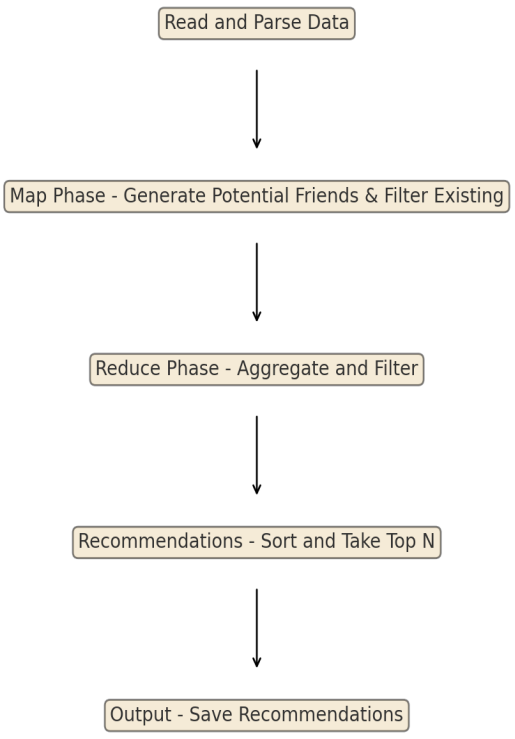
CS 6220 Data Mining | Homework 3

Yixuan Huang
https://github.com/yxhuang999/data-mining-neu

Pipeline Sketch:

This PySpark recommendation algorithm processes a social network dataset to generate friend recommendations based on mutual friends. It reads and parses user-friendship data, maps potential friend pairs while counting mutual friends, and reduces by aggregating these counts to filter out existing friendships. Recommendations are sorted by the number of mutual friends, taking the top N for each user. Finally, the recommendations are saved to a file, completing the pipeline from data input to actionable friend suggestions

Diagram:



Output:

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