CS 6350: Big Data Management and Analytics

Assignment 2

Names of students in your group:

1. Kushal Choudhary (kxc240000)

Number of free late days used: 0

Sources and References:

1. LiveJournal Social Network Dataset: https://an-ml.s3.us-west-1.amazonaws.com/soc-LiveJournal1Adj.txt
2. 20 Newsgroups Dataset: http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz
3. Apache Spark Documentation: https://spark.apache.org/docs/latest/rdd-programming-guide.html
4. Python Libraries Used: - PySpark (core implementation) - urllib, os, re, collections (data preprocessing only)

Part 1: Friend Recommendation

1. Code: https://github.com/Kushal3121/Friend-Recommendation/recommend_friends.py
2. Report: https://github.com/Kushal3121/Friend Recommendation/report.md
3. Dataset:
    a. LiveJournal Social Network (4.8M users, 68M edges)
    b. https://an-ml.s3.us-west-1.amazonaws.com/soc-LiveJournal1Adj.txt
4. Summary: Implemented MapReduce-based friend recommendation using mutual friends algorithm. Successfully generated recommendations for 10 randomly sampled users.

Part 2: Naïve Bayes Classification

1. Code: https://github.com/Kushal3121/Naive Bayes Classifier/naive_bayes.py
2. Report: https://github.com/Kushal3121/Naive-Bayes-Classifier/ALGORITHM_NAIVE_BAYES.md
3. Dataset:
    a. 20 Newsgroups (18,846 documents, 20 categories)
    b. http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz
4. Summary: Implemented Naive Bayes classifier from scratch using MapReduce with Laplace smoothing. Achieved 78.97% accuracy on 20-class text classification problem.