Homework #2: Discovering Association Rules via Spark

Due: October 18, Friday
100 points

Consider the movie lens data set: https://www.kaggle.com/shubhammehta21/movie-lens-small-latest-dataset/. Here we only consider the "ratings.csv" file which has 100,836 rows (ignore the header). We are only concerned with the first two columns: userId and movieId. Your task is to implement a Spark algorithm, assoc.py, for discovering association rules of the form: I → j, where I is an itemset and j is a single item (similar to what the text book discusses), from the dataset. Note that items here are movies and users are baskets.

Requirements:

- Your algorithm should first discover frequent itemsets with the specified threshold for support count.
- The discovery of frequent items should be done in parallel by following the SON algorithm and using mapPartitions() to process each chunk/partition of data by implementing an Apriori algorithm.
- You should make the chunk size small enough so that it can be loaded entirely into memory.
- As immediate results, your algorithm should also output the discovered frequent itemsets (i.e., movies frequently watched by many users).
- The discovering of association rules should be done in parallel and based on the discovered frequent itemsets. Note that we assume that the support count for I U {j} ≥ the support threshold.
- The confidence of the discovered association rules should meet or exceed the specified threshold.

Execution format:

spark-submit assoc.py ratings.csv <support threshold> <confidence threshold>

where the support threshold is an integer (for support count) and the confidence threshold is a value between 0 and 1.