Due Date
February 6, 2020

Assignment 2

CS 6665
Spring 2020

Association Rules

The action or practice of selling additional products or services to existing customers is called cross-selling. Giving product recommendation is one of the examples of cross-selling that are frequently used by online retailers. One simple method to give product recommendations is to recommend products that are frequently browsed together by the customers.

Suppose we want to recommend new products to the customer based on the products they have already browsed on the online website. Write a program using the A-priori algorithm to find products which are frequently browsed together. Fix the support to s = 100 (i.e. product pairs need to occur together at least 100 times to be considered frequent) and find itemsets of size 2 and 3.

Use the online browsing behavior dataset “browsing.txt” as the input file. Each line represents a browsing session of a customer. On each line, each string of 8 characters represents the id of an item browsed during that session. The items are separated by spaces.

Identify pairs of items (X, Y) such that the support of $\{X, Y\}$ is at least 100. For all such pairs, compute the confidence scores of the corresponding association rules: $X \Rightarrow Y$, $Y \Rightarrow X$. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the report. Break ties, if any, by lexicographically increasing order on the left hand side of the rule.

Identify triples of items (X, Y, Z) such that the support of $\{X, Y, Z\}$ is at least 100. For all such triples, compute the confidence scores of the corresponding association rules: $(X, Y) \Rightarrow Z$, $(X, Z) \Rightarrow Y$, $(Y, Z) \Rightarrow X$. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the report. Order the left-hand-side pair lexicographically and break ties, if any, by lexicographical order of the first then the second item in the pair.

Bonus (2pt)

Implement the A-priori algorithm on Spark.

What to submit

- Submit the source code
- Include the top 5 association rules for pairs and top 5 association rules for triples.

For a sanity check, the top two association rules for pairs are:

- DAI93865 → FRO40251 : 1.000000
- GRO85051 → FRO40251 : 0.999176