Lab 4

Association Rules Mining

CS429 - Introduction to Big Data Analysis

Lab Instructor: Nguyễn Đình Thảo

Table of Contents

Association Rules Mining Problem

Problem Statement

Given a list of transactions, each transaction contains a set of items. Find all association rules from such a list that satisfy:

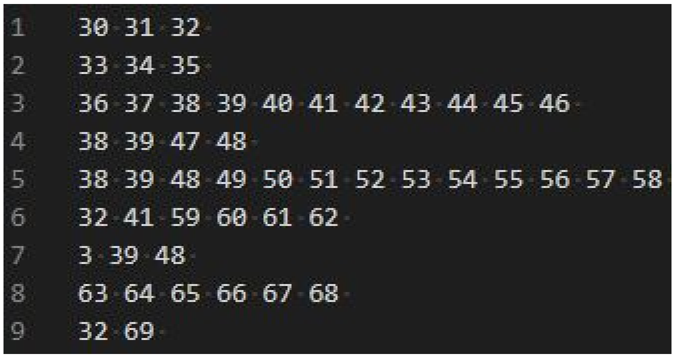
$$support \geq 2$$

$$confidence \geq 70\%$$

See section 6.1.1 and 6.1.3 from *Mining of Massive Datasets* for definition of Frequent Itemsets and Association Rule.

Input

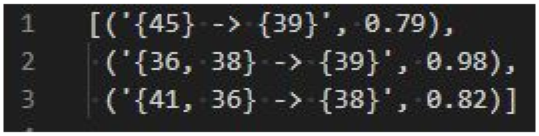
Sample data can be found in ***retail.dat***. Each line is one transaction, where items are separated by space.



```
30 31 32
33 34 35
36 37 38 39 40 41 42 43 44 45 46
38 39 47 48
38 39 48 49 50 51 52 53 54 55 56 57 58
32 41 59 60 61 62
3 39 48
63 64 65 66 67 68
32 69
```

Output

Output should be a list of association rules, each rule is a tuple of *(rule, rule_confidence)*.
For example:

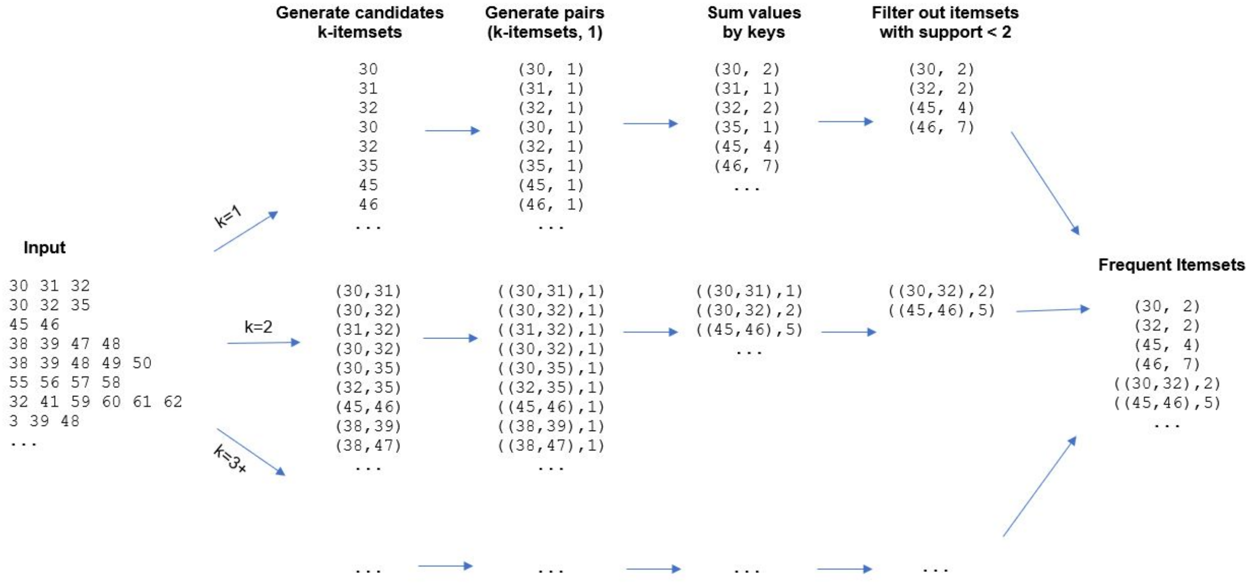


```
[('{45} -> {39}', 0.79),
 ('{36, 38} -> {39}', 0.98),
 ('{41, 36} -> {38}', 0.82)]
```

Pseudocode

A naive frequent itemsets generation

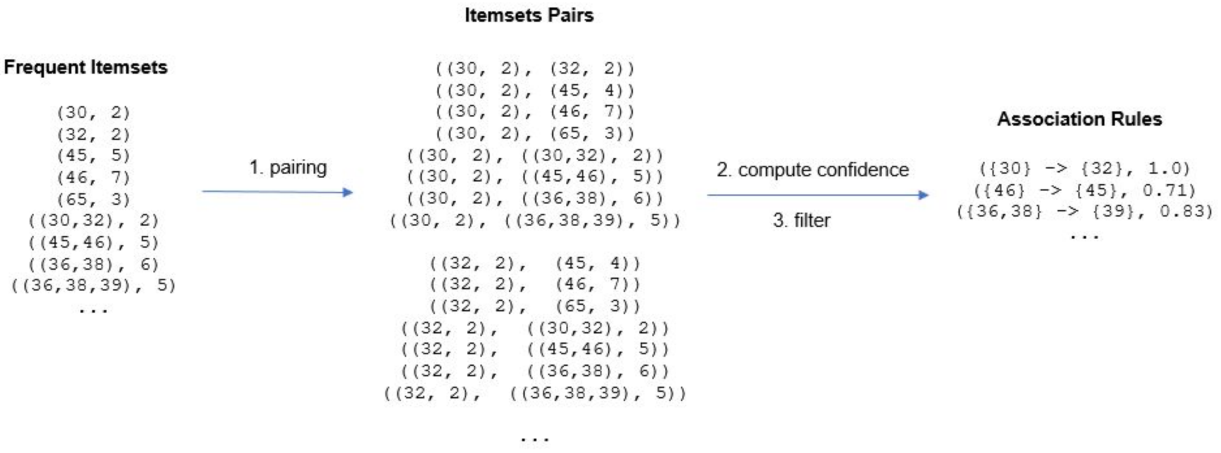
```
While new frequent itemset can be generated do:
    Start from $k = 1$:
        Generate candidate $k$-itemsets
        Generate pairs ($k$-itemsets, 1)
        Sum values in all pairs by keys
        Filter out itemsets with support < 2
Merge all $k$-itemsets to yield the frequent itemsets.
```



Frequent Itemsets generation

Association Rules mining

Given the frequent itemsets as result from the previous step, you need to compute the confidence for each association rule and then filter out rules with low confidence. One way to achieve this is depicted in the figure below.



Association rules mining demonstration

Lab Assignment

Implement the provided pseudocode for association rules mining in PySpark.
Submit your jupyter notebook with the naming format: **<your studentID>_lab4.ipynb**.