# Homework 2: Discovery of Frequent Itemsets and Association Rules

MINHA CHOI RALFS ZANGIS
minhac | zangis @kth.se

November 20, 2021

## Contents



## 1 Description of your solution

The final solution was made in Python programming language and incorporated the usage of libraries such as Python numpy, collections. In this task we have one data file ”T10I4D100K.dat” and one ”.py” file as name as ”apriori.py” including main script. The end product has been made available in a public repository (Accessible using the following link: Homework 2).

For the A-priori algorithm, the pipeline was built to create further candidates sets and frequent itemsets, and an algorithm was implemented to generate association rules.

## 2 Instructions how to build and to run

For the solution to run, the user is required to have python installed on their machine and have access to previously mentioned files. With the prerequisites fulfilled all that should be required is running the following command in the folder containing the files (no additional libraries should be needed for the execution):

```
$ python apriori.py
```

## 3 Results

In this task, we find the frequent itemsets with the data set containing sales transactions. During the testing, support and confidence threshold values are 700 & 1000 and 0.6, respectively.

Finding frequent itemset was conducted until there were no more frequent itemsets than the threshold support. As a result, we have found that

- **Subject = 700**
  476 singletone, 87 doubletone, 29 tripletone, 10 quadrupletone, 1 quintupletone

- **Subject = 1000**
  375 singletone, 9 doubletone, 1 tripletone

Figure 1 shows frequent itemsets but we skipped attaching singleton result.

```
Doubletone [9] :
[(('217', '346'), 1336), (('368', '829'), 1194), (('789', '829'), 1194), (('368', '682'), 1193), (('39', '825'), 1187), (('39', '704'), 1107),
(('704', '825'), 1102), (('227', '390'), 1049), (('390', '722'), 1042)]

Tripletone [1] :
[(('39', '704', '825'), 1035)]
```

Figure 1: doubletone and tripletone result where threshold subject=1000

Figure 2 shows generated association itemsets having larger confidence then 60 percentage.

```
('704',) -> ('825',) : 0.61%
('704',) -> ('39',) : 0.62%
('39', '704') -> ('825',) : 0.93%
('39', '825') -> ('704',) : 0.87%
('704', '825') -> ('39',) : 0.94%
```

Figure 2: Result of association itemsets where threshold subject=1000, confidence=0.6

Elapsed times for two tasks are that

- **Subject = 700**
  Finding frequent itemset : 113s
  Generating association itemset : 0.015s
- **Subject = 1000**
  Finding frequent itemset : 13.24s
  Generating association itemset : 0.0004s

# 4 Conclusion

In the conclusion, since the amount of pruning in the filter varies depending on the threshold support value, it seems important to set it as an appropriate value because the time required and the result of the item set differ.