# DAMAGE

## *Eclat*

Eclat was run with a bare minimum of custom settings to get a list of frequent itemsets

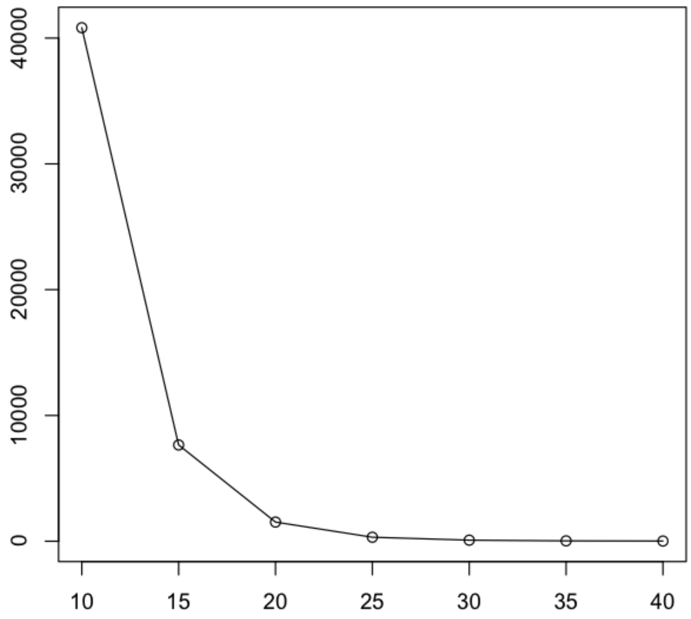
```
eclat -s10 -q-1 course-text.txt course10s.txt
```

-s denotes the threshold (10 in this case)
-q denotes sorting in a descending order.

## *Results*

Below is the number of frequent itemsets per a specific threshold. As we can see the number of frequent itemsets grow exponentially, as we weaken the threshold

| Threshold | 40 | 35 | 30 | 25 | 20 | 15 | 10 |
|---|---|---|---|---|---|---|---|
| # of freq.itemsets | 15 | 28 | 84 | 317 | 1522 | 7653 | 40823 |



## *Descriptive statistics*

The results of the frequent itemset detection with eclat do not give us to much interpretation possibilities, therefore we decided to have a closer look at the basic descriptive statistics. Hoping that a better understanding of the data itself will benefit our choice of parameters used with eclat.

Courses in total: 99

Courses per transaction:

| Min | 1st Qu. | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|
| 1 | 4 | 11 | 12.96 | 22 | 40 |

Number of transactions: 2401

The most frequent courses will appear more often in itemsets than courses with average frequency. This does not necessary mean, that those courses can be better associated to other courses. It is more likely, that those courses are more frequent, because they are basic courses, or even mandatory courses.
If a course is mandatory, the students choice to take this course was limited, therefore itemset containing mandatory course do not reflect the use of the academic freedom of the students and have to be handled carefully.
Also association rules containing frequent courses, will more likely be determined on the less frequent courses belonging to the association rule.

To conclude, the relevance of a not appearance of a frequent course in certain association rules might be higher, than the actual appearance.

**Top 20 of the most frequent courses:**

| | |
|---|---|
| Programming Project: | 4.81% |
| Computer Organization: | 3.89% |
| Introduction To Unix: | 3.47% |
| Information Systems: | 3.44% |
| Data Structures: | 3.44% |

Concurrent Systems: 3.43%
Data Structures Project: 3.41%
Data Communications: 3.34%
Introduction To Computing: 3.20%
Programming In C: 2.96%
Programming (pascal): 2.75%
Information Systems Project: 2.45%
Scientific Writing: 2.43%
Database Systems I: 2.40%
Software Engineering: 2.36%
Models For Programming And Computing : 2.24%
Introduction To Application Design: 2.17%
Programming In Java: 2.10%
Introduction To Databases: 2.09%
Introduction To Programming: 2.09%

Knowledge of the courses with the lowest frequency will help us to detect itemsets and associations rule, who simply appear by chance or who are highly reliable. Since they appear with a low frequency, their 1-itemset support is already low. If those courses are still associated with other courses, this most likely is coincidental, or in special cases, can hint to a high reliability.
However, we have to keep in mind, that some of the courses might not be offered anymore, have only been offered for a short time or were replaced with similar courses. If this is the case, their frequency is naturally lower than the average, since they only could appear in a subset of the transaction, so choices of courses of a student. If, we want to get a closer look at those courses, we should run analysis concerning only the subset of student study plans, which actually had the possibility to include those. Also, information about those courses, which might not be in the given data itself, will help to interrogate their association rules.

We can conclude, that a frequent appearance of those courses in our item sets stems from a to low chosen threshold.

**Bottom 20 courses**

Tietojenkasittelytieteen Approbatur Pääaineoppimäärä: 0.17%
Electronic Commerce And Internet: 0.15%
Management Of Research Data: 0.14%
Algoritmisen Tietojenkäsittelyn Perusteet: 0.13%
Computing Methodologies: 0.13%
Database Modelling: 0.13%
An Introduction To Specification And Verification: 0.13%
Philosophy Of Artificial Intelligence: 0.12%

Spatial Information Systems: 0.12%
Verification And Derivation Of Algorithms: 0.12%
Robotiikka: 0.12%
Symbolic Programming: 0.12%
Introduction To Computers: 0.12%
Distributed Systems: 0.11%
Geometric Methods: 0.11%
Seminaari: Adaptiiviset Oppimisymparistöt : 0.11%
Machine Learning: 0.11%
Computer Architectures: 0.10%
Information Extraction In Text: 0.10%
Three Concepts: Information: 0.10%