# Data Mining
# Homework Assignment #2

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

February 20, 2014

Table 1: Example of the transaction data set

| CustomerID | TransactionID | BasketContent |
|---|---|---|
| 1 | 1234 | {Aspirin, Panadol} |
| 1 | 4234 | {Aspirin, Sudafed} |
| 2 | 9373 | {Tylenol, Cepacol} |
| 2 | 9843 | {Aspirin, Vitamin C, Sudafed} |
| 3 | 2941 | {Tylenol, Cepacol} |
| 3 | 2753 | {Aspirin, Cepacol} |
| 4 | 9643 | {Aspirin, Vitamin C} |
| 4 | 9691 | {Aspirin, Ibuprofen, Panadol} |
| 5 | 5313 | {Panadol, Vitamin C} |
| 5 | 1003 | {Tylenol, Cepacol, Ibuprofen} |
| 6 | 5636 | {Tylenol, Panadol, Cepacol} |
| 6 | 3478 | {Panadol, Sudafed, Ibuprofen} |

## Task 1

a. For the data in 1 compute the support and support count for itemsets {Aspirin}, {Tylenol, Cepacol}, {Aspirin, Ibuprofen, Panadol} by treating each transaction ID as a market basket.

b. Compute the confidence for the following association rules: {Aspirin, Vitamin C $\rightarrow$ Sudafed}, {Aspirin $\rightarrow$ Vitamin C}, {Vitamin C $\rightarrow$ Aspirin}. Why the results for last two rules are different?

c. List all the frequent itemsets under the support count threshold $s_{min} = 3$.

d. What does the anti-monotonicity property of a support imply? Give an example using the above data set.

## Task 2

Write down all the steps of Apriori algorithm on the data set 1 under the support count threshold $s_{min} > 3$. How many steps of Apriori algorithm you needed to perform? Draw a diagram showing all possible combinations of the items (e.g. lecture slide number 68). Mark all maximal, closed and infrequent items on this diagram.

## Task 3

Construct an FP-tree using the same data set 1 (use support count threshold $s_{min} > 3$) . Explain all the steps of the tree construction and draw a resulting tree. Based on this tree answer the questions: how many transactions contain {Aspirin} and {Cepacol}? How many transactions were made in total?

## Task 4

Simulate frequent pattern enumeration based on the FP-tree constructed in the previous exercise. Report all the frequent patterns.

## Task 5

In this task we will get familiar with the statistical computing language R. Install it. We suggest you to download also the IDE that will make your life much easier: R studio. Once you are set up, take a look at the introduction of R from the CRAN page (Manuals → An Introduction to R) or just google any basic tutorial. R is an open source and has a very powerful community with plenty of tutorials and websites. Once you feel more comfortable with it, go through the following tutorial, run it, check and report the results, describe and interpret them:

---

```
#install necessary packages (run only once)
install.packages("arules")
install.packages("arulesViz")

#load datafrom the url
data_url =
    url("https://courses.cs.ut.ee/MTAT.03.183/2014_spring/uploads/Main/titanic.txt")
titanic = read.table(data_url, sep =',', header = TRUE)

#observe the data
##first 6 observations
head(titanic)
#types of features
str(titanic)
```

```r
#dimensionality of the data
dim(titanic)

#load package for frequent set mining
library(arules)
#help with apriori
?apriori
#run apriori algorithm with default settings
rules = apriori(titanic)
#inspection of the result
inspect(rules)

#now let us assume, we want to see only those rules that have
    rhs as survived:
rules = apriori(titanic,appearance = list(rhs=c("Survived=No",
    "Survived=Yes"),default="lhs"))
inspect(rules)

#let us relax the default settings for the rules we are
    looking for
rules = apriori(titanic,parameter = list(minlen=2, supp=0.05,
    conf=0.8),appearance = list(rhs=c("Survived=No",
    "Survived=Yes"),default="lhs"))

#visualization
library(arulesViz)
plot(rules, method="graph", control=list(type="items"))
```

## Task 6 (2pt)

Use the drug dataset 1 (or simulate a similar to it) and repeat the analysis in Task 5 using R. Report the results.