Submission sheet
Assignment II

This document contains questions to help you reflect about the operations applied to the data during this assignment. You have to fill it in and submit it on Studium.

TASK 1: warm up

Number of records 9816 (after removing duplicates)

Number of attributes (excluding User and Query) 500

TASK 2: frequent itemsets

What is the most frequent item (i.e., single keyword)? "of"

In how many queries does it occur: 955 -> without removing duplicates (927 with removing duplicates)

Which support value corresponds to a support count of 100 records? 100 / 9816 = 0.0101874491

How many frequent itemsets have you found with this min-support? 28

What is the maximum size of your frequent itemsets? 2

TASK 3: impact of the support parameter

Number of frequent itemsets for some selected values of support:

Mininum support	Number of Frequent Itemsets
0.001	855
0.004	140
0.007	55
0.01	27

Which value of min-support leads to the discovery of about 150 itemsets? 0.004 -> leads to 140 itemsets

TASK 4: rule generation

How many rules have you identified with min-confidence .8? 4

Indicate a high-confidence rule X -> Y where Y -> X has lower confidence.

York → new

Why is the confidence of Y -> X lower than the confidence of X -> Y?

Because the word "York" is amost exclusively used with "New York" whereas the word "new" can be used in many more contexts other than "New York". For example "new car", "new job"...

TASK 5: impact of confidence

Number of rules for some selected values of minimum confidence:

Mininum confidence	Number of Rules
0.1	18
0.3	14
0.5	7
0.7	5
0.9	1

TASK 6: rule interpretation

What are the top-3 interesting rules, and what is their rank w.r.t. confidence and lift?

RULE			RANK (Conf)	RANK (Lift)
humane	→	society	1	377.54
cheat	→	codes	0.6667	261.76
social	→	security	0.5556	259.68

What are the top-3 unexpected rules, and what is their rank w.r.t. confidence and lift?

RULE			RANK (Conf)	RANK (Lift)
2	→	sims	0.2	151.02
st.	→	louis	0.2632	117.42
2006	→	april	0.1149	80.59