

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 (after removing duplicates)

Number of attributes (excluding User and Query)

TASK 2: frequent itemsets

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

In how many queries does it occur:

Which support value corresponds to a support count of 100 records?

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

What is the maximum size of your frequent itemsets?

TASK 3: impact of the support parameter

Number of frequent itemsets for some selected values of support:

Minimum support	Number of Frequent Itemsets
<input type="text" value="0.001"/>	<input type="text" value="855"/>
<input type="text" value="0.004"/>	<input type="text" value="140"/>
<input type="text" value="0.007"/>	<input type="text" value="55"/>
<input type="text" value="0.01"/>	<input type="text" value="27"/>

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

TASK 4: rule generation

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

Indicate a high-confidence rule $X \rightarrow Y$ where $Y \rightarrow X$ has lower confidence.

\rightarrow

Why is the confidence of $Y \rightarrow X$ lower than the confidence of $X \rightarrow Y$?

Because the word "York" is almost 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:

Minimum 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	2	1
diego	→ san	4	2
credit, federal	→ union	5	3

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

RULE		RANK (Conf)	RANK (Lift)
american	→ african	177	66
2006	→ april	178	67
center	→ medical	176	65