Data Mining - Blatt 01

Thomas, Manuel, Marius

October 31, 2017

1 Nr 1

Je höher der Support-Threshold, je weniger Kombinationen in der Ausgabe. Das liegt daran, dass der Support-Threshold angibt, wie hoch die Wahrscheinlichkeit einer Kombination sein muss.

2 Nr 2

Histogramme sind im Ordner Histogramme_Nr2.

3 Nr 3

Die Itemsets von dm2 kommen weniger häufig vor als die vom dm3. Bim dm2 gibt es ein 1-Itemsets, welches in 90% der Fälle vorkommt. dm3 hat relativ viele Kombinationen von Itemsets, deswegen zeigt das Histogramm auch relative viele Häuffungen bei den einzelnen Support-Werten.

4 Nr 4

Datensatz	Min-Support	Runtime
dm1	0.4	1712 function calls in 0.028 seconds
dm1	0.5	980 function calls in 0.012 seconds
dm1	0.6	831 function calls in 0.011 seconds
dm1	0.7	732 function calls in 0.012 seconds
dm1	0.8	668 function calls in 0.009 seconds
dm1	0.9	562 function calls in 0.007 seconds
dm2	0.4	757 function calls in 0.010 seconds
dm2	0.5	490 function calls in 0.009 seconds
dm2	0.6	445 function calls in 0.007 seconds
dm2	0.7	445 function calls in 0.007 seconds
dm2	0.8	445 function calls in 0.007 seconds
dm2	0.9	445 function calls in 0.008 seconds
dm3	0.4	26262 function calls in 0.156 seconds
dm3	0.5	7805 function calls in 0.055 seconds
dm3	0.6	4143 function calls in 0.030 seconds
dm3	0.7	2025 function calls in 0.017 seconds
dm3	0.8	1068 function calls in 0.013 seconds
dm3	0.9	664 function calls in 0.010 seconds
movielines	0.4	1932402 function calls in 13.593 seconds
movielines	0.5	1932391 function calls in 13.791 seconds
movielines	0.6	1932391 function calls in 13.480 seconds
movielines	0.7	1883095 function calls in 13.674 seconds
movielines	0.8	1883084 function calls in 13.977 seconds
movielines	0.9	1883084 function calls in 13.460 seconds

Je niedriger der Min-Support ist, desto höher wird die Runtime. Das liegt daran, dass der Algorithmus dann viel mehr Möglichkeiten durchgehen muss als wenn der Min-Support größer ist. Bei einem hohem Min-Support schneidet der Algorithmus gleich mehrer Möglichkeiten weg und braucht deswegen weniger Functionsaufrufe und somit eine geringere Runtime.

5 Nr 5

Die Ausgabe für dm3 mit einem Min-Support von 0.7 liegt im Ordner abgabe_nr5. Die csv Datei hat in jeder Zeile ein frequent itemset. Die Labels sind dabei mit einem ;getrennt. Die Labels sind die Spalten der Data-CSV.