T-61.5060 Algorithmic methods in data mining
Exercises 3, September 27, 2007

No lecture on October 1; no exercises on October 4.

1. Find the frequent sets from the data set below using (absolute) frequency threshold 3. List also a few association rules.

A	B	C	D	E
1	1	0	1	0
0	1	1	1	0
0	0	1	0	0
1	0	1	1	1
0	1	1	0	1
0	1	1	0	1
1	1	1	0	1
1	0	1	1	1
0	1	1	1	1
1	0	1	1	1

2. What data structures are needed for the implementation of the levelwise algorithm for finding frequent sets? Assume you for some reason would not be able to use an existing implementation for finding frequent sets, but would have to write a new one from scratch. What language and what structures would you use if the program would be used a few times on datasets containing a few thousand rows, or many times on multimillion row datasets?

3. Compute the frequent set collection on the data sets abstracts10000.txt for some sensible values of the frequency threshold. Try to interpret some of the resulting sets. How does the size of the collection grow as a function of the threshold?

4. Let D be a dataset, and form another dataset D' by adding k new columns to D; the new columns are filled with random bits (probability of a 1 is 0.5, independently from all other positions in the data). What can be said about the frequent sets of D and D'?

5. Let D be a dataset, and form another dataset D' by independently flipping each 1 in D to 0 with probability q. What can be said about the frequent sets of D and D'?