T-61.5060 EXERCISE 6/2005

In T3 on 23 November 2005 at 12 o'clock.

No exercise session on 16 November 2005 and no lecture 17 November 2005 due to a travel.

To pass the course you must pass the examination and complete the exercise work. The exercise work must be submitted by 16 January 2006, see `http://www.cis.hut.fi/Opinnot/T-61.5060/2005/harjoitustyo.shtml` for instructions.

1. Generate a random dataset with 10000 rows and 100 variables and compute the frequencies of all pairs of variables. Take samples of size K, where $K \in \{100, 500, 1000\}$, and look what is the maximal absolute error in the frequencies. Do the results correspond to the bounds given in the lectures (or in [1])?

2. Suppose you have a the 0-1 dataset where the rows correspond to documents and the columns correspond to words. Suppose there are 1000000 documents and 100000 words, and assume you would want to reduce the dimension of the data somehow. What would you do?

3. Suppose you have two tables, R with attributes A and B and S with attributes B and C. Consider the join T of the tables, i.e., the collection of all triplets (a, b, c) such that (a, b) is a row of R and (b, c) is a row of S. How would you form a sample from T without computing the whole set T?

References

[1] Gabor Lugosi. Concentration-of-measure inequalities. Lecture Notes, 7 March 2005. `http://www.econ.upf.es/~lugosi/anu.pdf`