T-61.5060 Algorithmic methods in data mining

Exercises 2, September 20, 2007

1. The file abstracts10000.txt available from the web site of the course contains information about the words occurring in 10000 NSF grant abstracts. Each row corresponds to a document, and the numbers in a row refer to the words occurring in that document. (The file words.txt gives the correspondence between words and numbers.)

   Count the number of occurrences for each word, and use some tools to model the distribution. That is, try to find a function f such that $f(k)$ indicates how many times the kth most common word occurs.

2. Using the same data set as in the previous exercise, take a random sample of 1000 documents and check how good estimates for the frequencies of words you get.

3. Which pair of words occurs most frequently in the dataset abstracts10000.txt? How did you determine this?

4. Chernoff bounds are used to bound the tails of binomial distributions. Other ways of approximating the binomial distribution include the normal approximation and the Poisson approximation. What are these approximations and when can they be applied?

5. Go through the details of the proof of the performance guarantee of the count-min data structure. Work out some examples on what the bound actually implies.

6. Experiment with a frequent set discovery algorithm; for code, see, e.g., the web page of Bart Goethals.