Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Marjan Celikik

Search Engines WS 09/10

http://ad.informatik.uni-freiburg.de/teaching

Exercise Sheet 6

complete until Tuesday, December 1st

Exercise 1

Consider your collection from Exercise Sheet 1. From the vocabulary of that collection pick a variety of ten three-letter prefixes. For each of these prefixes compute the following. (1) The total size of the inverted lists of all words matching the prefix. (2) The size of the largest inverted list of a word matching the prefix. (3) The ratio between the two. Write a couple of sentences about the meaning of these ratios.

Exercise 2

If you haven't already done it for one of the earlier exercise sheets, implement a k-way merge of k inverted lists, as discussed in the lecture. Then compute the following for each of your ten prefixes from Exercise 1. (1) The cost of merging the inverted lists of all words matching the prefix. (2) The cost of scanning the largest inverted list of a word matching the prefix. (3) The cost of scanning the inverted lists of all words in the collection. (4) Compute the ratio between (1) and (3) and between (2) and (3). Write a couple of sentences about the meaning of these ratios.

Exercise 3

Show that the following two ways of generating an inverted list of expected size m with doc ids from the range $1..n$ lead to exactly the same entropy. (1) Pick each of the n doc ids with probability $p = m/n$, independently from each other. (2) Pick each of the m gaps with the appropriate probability distribution as discussed at the beginning of Lecture 5, independently from each other.

Hint: Proceed as follows. (0) Recall that the entropy of a random variable X with $\Pr(X = i) = p_i$ is defined as $\sum_i p_i \cdot \log_2(1/p_i)$. (1) Compute the entropy H_1 of the random variable I that indicates whether a fixed doc id is in the inverted list or not, that is, $\Pr(I = 1) = p$ and $\Pr(I = 0) = 1 - p$. (2) Compute the entropy H_2 of the random variable G for a fixed gap, that is, $\Pr(G = k) = (1 - p)^k \cdot p$. (3) Prove that $n \cdot H_1 = m \cdot H_2$. Surprising, isn't it?

Exercise 4

Prove the following inequality, that was used in the estimation of the empirical entropy of the HYB index in the lecture: for all $x > 0$, it holds that $(1 + 1/x) \cdot \ln(1 + x) \leq 1 + x/2$.

Hint: the standard approach of setting the derivative of the difference between the left-hand side *and the right-hand side to zero will not work, because that derivative is not defined at 0. Whatever you do, come up with a rigorous proof, not some hand-waving argument, or proof via gnuplot, or something like that.*