58093 String Processing Algorithms (Autumn 2011)
Course Exam, 15 December 2011 at 16-19 Lecturer: Juha Kärkkäinen

Please write on each sheet: your name, student number or identity number, signature, course name, exam date and sheet number. You can answer in English, Finnish or Swedish.

1. [*4+4+4 points*] Each of the following pairs of concepts are somehow connected. Describe the main connecting factors or commonalities as well as the main separating factors or differences.

   (a) String quicksort and string mergesort.

   (b) Shift–Or algorithm and BNDM algorithm.

   (c) LLCP and RLCP array in string binary search and LCP array augmenting suffix array.

   A few lines for each part is sufficient.

2. [*6+8 points*] A q-gram of a string is its factor of length q. Let $G_q(A, B)$ denote the number q-grams shared by the strings A and B.

   For example, for $A = \texttt{varaurat}$ the 2-grams are `va`, `ar`, `ra`, `au`, `ur`, `ra` and `at`, and for $B = \texttt{ararat}$ they are `ar`, `ra`, `ar`, `ra` and `at`. The shared 2-grams are `ra` twice, `ar` and `at`, and thus $G_q(A, B) = 4$.

   (a) Show that if $ed(A, B) \leq k$, then $G_q(A, B) \geq |A| - q + 1 - kq$.

   (b) Design a filtering algorithm for approximate string matching based on the result of (a)-part.

3. [*6+6 points*] Let $\Sigma = \{\texttt{a}, \texttt{b}\}$ be the alphabet. For any integers $k \geq 1$ and $m \geq k$, describe a set of $n = 2^k$ strings of length m such that the number of nodes in the (uncompact) trie for the set is

   (a) as large as possible

   (b) as small as possible.

   What is the number of nodes in each case? Note that all the strings in the set must be different.

4. [*12 points*] Let T be a string of length n over an alphabet Σ of constant size. Describe an algorithm that finds the *longest* string over the alphabet Σ that occurs *exactly* k times in T. The time complexity should be $\mathcal{O}(n)$.