58093 String Processing Algorithms (Autumn 2013)
Exercises 2 (5 November)

Solve the following problems before the exercise session and be prepared to present your solutions at the session.

1. Outline algorithms that find the most frequent symbol in a given string

   (a) for ordered alphabet, and

   (b) for integer alphabet.

   The algorithms should be as fast as possible. What are their time complexities?

2. Let $\mathcal{R}$ be a set of n random strings from Σ^k for some $k > \log_\sigma n$.
   Show that $\Sigma lcp(\mathcal{R}) = \mathcal{O}(n \log_\sigma n)$ on average.

3. Show that (see page 23 on the lectures):

   (a) For $i \in [2..n]$, $LCP_{\mathcal{R}}[i] = lcp(S_i, S_{i-1})$.

   (b) $\Sigma LCP_\pi(\mathcal{R}) = \Sigma LCP(\mathcal{R})$.

4. Let $\mathcal{R} = \{\texttt{manne}, \texttt{manu}, \texttt{minna}, \texttt{salla}, \texttt{saul}, \texttt{sauli}, \texttt{vihtori}\}$.

   (a) Give the compact trie of $\mathcal{R}$.

   (b) Give the balanced compact ternary trie of $\mathcal{R}$.

5. What is the time complexity of prefix queries for

   (a) trie with constant alphabet

   (b) compact trie with constant alphabet

   (c) compact trie with ordered alphabet and binary tree implementation of the child function

   (d) balanced compact ternary trie?

   The queries should return the resulting strings as a list of pointers or other identifiers rather than the full strings.

6. Complete the proof of Theorem 1.12 by showing the following result:

   Let $n_1, n_2, \ldots, n_d$ be positive integers, and let $n = \sum_{i=1}^{d} n_i$. Then

   $$\sum_{i=1}^{d} n_i \log n_i \geq n \log \frac{n}{d}$$

   Hint: Look up Jensen's inequality.