

58093 String Processing Algorithms (Autumn 2012)

Exercises 2 (8 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. Prove Lemma 1.11.
3. Let \mathcal{R} be a set of n random strings from Σ^k for some $k > \log_\sigma n$. Show that $lcp(\mathcal{R}) = \mathcal{O}(n \log_\sigma n)$ on average.
4. Let $\mathcal{R} = \{\text{manne, manu, minna, salla, saul, sauli, vihtori}\}$.
 - (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 1.3 by showing the following result:

Let n_1, n_2, \dots, 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.