

58093 String Processing Algorithms (Autumn 2016)

Exercises 2 (Tuesday, November 8)

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) in general alphabet model, and
- (b) in integer alphabet model.

The algorithms should be as fast as possible. What are their (worst case) time complexities? Consider also the case where $\sigma \gg n$.

2. 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} .

3. What is the time complexity of prefix queries for

- (a) trie in constant alphabet model
- (b) compact trie in constant alphabet model
- (c) compact trie in general alphabet model using a 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.

4. Prove

- (a) Lemma 1.14: For $i \in [2..n]$, $LCP_{\mathcal{R}}[i] = lcp(S_i, \{S_1, \dots, S_{i-1}\})$.
- (b) Lemma 1.15: $\Sigma LCP(\mathcal{R}) \leq \Sigma lcp(\mathcal{R}) \leq 2 \cdot \Sigma LCP(\mathcal{R})$.

5. Show how to construct the compact trie for a set \mathcal{R} in $\mathcal{O}(|\mathcal{R}|)$ time (rather than $\mathcal{O}(|\mathcal{R}|^2)$ time) given the string set \mathcal{R} in lexicographical order and the LCP array $LCP_{\mathcal{R}}$.