Technische Universität München
Fakultät für Informatik
Lehrstuhl für Theoretische Informatik
Prof. Dr. Susanne Albers
Dr. Arindam Khan
Maximilian Janke
Luisa Peter

Winter Semester 2018/19
Problem Set 7
December 11, 2018

Advanced Algorithms

Due December 18, 2018 at 10:00

Note: You are welcome to submit in groups of two. If you wish to submit individually, exercises 1 and 2 are to be solved.

Exercise 1 (McCreight's algorithm – 10 points)

Construct the suffix tree T for $\sigma = mississippi\$$ using McCreight's algorithm. Draw all trees T_i and don't forget to insert the suffix links. Remember that in T_i all internal nodes have a suffix link, except for the internal node possibly inserted into T_i in iteration i.

Exercise 2 (Generalized Suffix Trees – 10 points)

The generalized suffix tree for a set of strings $\sigma_1, \sigma_2, \ldots, \sigma_k$ of total length n over alphabet Σ is the suffix tree of all suffixes of the strings $\sigma_1\$_1$, $\sigma_2\$_2$,..., $\sigma_k\$_k$. The $\$_1, \ldots, \$_k$ are pairwise distinct characters not in Σ. The label of each leaf in this generalized suffix tree is a pair (i, j) indicating that the path corresponds to suffix j of string σ_i.

1. Explain how to construct the generalized suffix tree for a set of strings.
   Hint: Use a suffix tree construction algorithm for a single string.

2. Draw the generalized suffix tree for the strings $\sigma_1 = acba$ and $\sigma_2 = cbaac$ following your explanations from 1.

3. Explain how to compute the length of a longest common substring of two given strings σ_1 and σ_2.

Exercise 3 (Applications of Suffix Trees – 10 points)

1. In molecular biology an RNA sequence is a string over the alphabet $\Sigma = \{A, C, G, U\}$ and an RNA molecule is called circular if it forms a closed loop. Hence, a circular RNA sequence has no natural starting point. In order to list circular RNA sequences in a database, they need to be displayed as linear strings in canonical form. Choosing the lexicographically smallest string among all possible linear strings is a natural choice for the canonical form.

   Describe an algorithm that utilizes a suffix tree to find the lexicographically smallest string representing a given circular RNA sequence. Analyze the running time of the algorithm.

2. Given some string σ of length n over an alphabet Σ. Describe an algorithm that finds a longest substring of σ that is a palindrome.
   Hint: You can use the generalized suffix trees introduced in exercise 2.

Exercise 4 (Ukkonen's Algorithm – 10 points)

Use Ukkonen's algorithm to construct the suffix tree for $t = remember\$$. Draw all implicit suffix trees T_i and write down which suffix extension rules you applied.