CSMTSP Text Searching and Processing

1. **(a)** Give all the periods and borders of the string

$$x = \texttt{abaababaabaababaababa}.$$

[10 marks]

(b) The table *Border* of a string x contains the length of the borders of the prefixes of the string x. Compute the table *Border* related to the word x of Q. **1(a)**. *[10 marks]*

(c) Describe in English or in pseudo-code how to compute the border of the string ua, where $u \in \Sigma^*, a \in \Sigma$, and Σ is the underlying alphabet, if you already know all the borders of the prefixes of u. *[20 marks]*

(d) Describe the criterion used on the table *Border* of x to find if some prefix of x is a square. (A square is a word of the form vv where $v \in \Sigma^*$ and is non-empty). *[10 marks]*

2. **(a)** Design the Aho-Corasick (AC) dictionary matching automaton over the alphabet $\Sigma = \{\texttt{a}, \texttt{b}\}$ for the following set of keywords:

$$\texttt{ababa},\ \texttt{bab},\ \texttt{bb}.$$

[20 marks]

(b) Describe in English or in pseudo-code the *Next_State* procedure used during the search for the occurrences of the keywords in a text. *[20 marks]*

(c) How would you implement a node of the AC automaton? *[10 marks]*

3. **(a)** Give the trie, without suffix links, of all the suffixes of the word `ababba`. *[10 marks]*

(b) Give the suffix tree, with suffix links, of the word `ababba`. *[20 marks]*

(c) Give the suffix automaton of the word `ababba`. *[10 marks]*

(d) How would you find the number of occurrences of a given substring of a text using its suffix tree? *[10 marks]*

4. **(a)** Define the k-differences approximate pattern matching problem. How would you initialize the dynamic programming matrix for such a problem? *[10 marks]*

(b) Let DP be the k-differences dynamic programming matrix of two strings x and y of lengths n and m respectively. Give the relation to compute $DP[i, j]$ for $0 < i \leq n, 0 < j \leq m$ where unit costs are applied for each operation. Give the DP matrix for the two strings $x =$ `abaabaabca` and $y =$ `baaca` with $k = 1$. *[15 marks]*

(c) Outline the trace-back strategy for locating the starting positions of the occurrences of a pattern in a text. *[15 marks]*

(d) Give the relation to compute $DP[i, j]$ as in Q. **4(b)** with weighted costs. *[10 marks]*

5. Consider a list of strings $L = (y_1, y_2, \ldots, y_k)$, in lexicographic order: $y_1 \leq y_2 \leq \ldots \leq y_k$. All of the strings have the same length n, and the list is to be searched for a target string x, also of length n.

(a) What is the asymptotic cost of a binary search for x in the list L if no extra information on the strings $y_1, \ldots, y_k$ is known? Give a "worst-case" example to illustrate your answer. *[15 marks]*

(b) What is the time complexity of the problem stated in Q. **5(a)** if the LCP (Longest Common Prefix) information is known? *[5 marks]*

(c) How many longest common prefixes of $y_1 \cdots y_k$ need to be preprocessed to run a binary search of the previous question Q. **5(b)**? *[10 marks]*

(d) Give the suffix array of the string `ababba`. *[10 marks]*

(e) What is the time complexity of searching for a pattern x in a text y, given its suffix array? *[10 marks]*

Final Page