

Special course in Computer Science:

Advanced Text Algorithms

Exercise set 1

Due: 8.11.2018

1. Compute the number of different

(a) prefixes and suffixes

(b) factors, and

(c) subsequences

of a given string of length n .

2. A circular string of length n is a string in which character n is considered to precede character 1. (Bacterial and mitochondrial DNA is typically circular.) Assume you have a subroutine (e.g., $BM(pat, text)$ or $KMP(pat, text)$) which checks whether a given linear string (pat) can be found within a linear text. Provide an algorithm that applies this subroutine to determine whether a linear string (par) is a substring of a circular string ($text^{circ}$). Analyse the algorithm's complexity with respect to the size of the pattern, the size of the text (its linear size, excluding its circular shift), and the complexity $P(m, n)$ of the linear matching subroutine

3. Construct both the table of borders and the table of strong borders for the pattern "maamamma".

4. Use the tables computed at 3. to simulate the MP algorithm for the text "jo_hommaamme_maamamman"

5. Same as above for the KMP algorithm.

6. Given a word w of length m , we define the function $R(x)$ for all characters x in the alphabet as follows: $R(x)=0$ if the character x does not occur in w and $R(x)=\max\{i < m \mid w[i]=x\}$ otherwise. Give an algorithm that computes the function $R(x)$ in $O(m)$ time.

7. Simulate Boyer-Moore matching to locate occurrences of pattern "maamamma" in text "jo_hommaamme_maamamman"

(a) applying the bad character shift rule alone

(b) applying the good suffix shift rule alone

(c) selecting the maximum shift given by the bad character rule and the good suffix rule.

(Note that in the lecture slides for the case of the bad character rule the mismatch is between $pat[i]$ and $text[i+j]$, while in the case of the good suffix rule the mismatch is between $pat[i-1]$ and $text[i+j-1]$)

8. (a) Find an example (i.e., pattern to be searched and text to search in) such that when using the "Bad Character Rule" alone within the Boyer-Moore matching algorithm it would generate $O(mn)$ comparisons.

(b) Same request for using only the "Good Suffix Rule"