National Institute of Technology Calicut
Department of Computer Science and Engineering
Third Semester B. Tech.(CSE)-Monsoon 2024
CS2091E Data Structures Laboratory
Assignment 8

Submission deadline (on or before): 11:59 PM, 04/11/2024

Policies for Submission and Evaluation:

- You must submit all the solutions of this assignment following the below-mentioned guidelines in the Eduserver course page, on or before the submission deadline.
- Ensure that your programs will compile and execute using GCC compiler without errors. The programs should be compiled and executed in the SSL/NSL.
- During the evaluation, failure to execute programs without compilation errors may lead to zero marks for that evaluation.
- Your submission will also be tested for plagiarism, by automated tools. In case your code fails to pass the test, you will be straightaway awarded zero marks for this assignment and considered by the examiner for awarding an F grade in the course. Detection of ANY malpractice related to the lab course can lead to awarding an F grade in the course.

Naming Conventions for Submission

- Submit a single ZIP (.zip) file (do not submit in any other archived formats like .rar, .tar, .gz). The name of this file must be

  `ASSG<NUMBER>_<ROLLNO>_<FIRST-NAME>.zip`

  (Example: $ASSG1\_BxxyyyyCS\_LAXMAN.zip$). DO NOT add any other files (like temporary files, input files, etc.) except your source code, into the zip archive.

- The source codes must be named as

  `ASSG<NUMBER>_<ROLLNO>_<FIRST-NAME>_<PROGRAM-NUMBER>.c`

  (For example $ASSG1\_BxxyyyyCS\_LAXMAN\_1.c$). If you do not conform to the above naming conventions, your submission might not be recognized by our automated tools and hence will lead to a score of 0 marks for the submission. So, make sure that you follow the naming conventions.

Standard of Conduct

- Violation of academic integrity will be severely penalized. Each student is expected to adhere to high standards of ethical conduct, especially those related to cheating and plagiarism. Any submitted work MUST BE an individual effort. Any academic dishonesty will result in zero marks in the corresponding exam or evaluation and will be reported to the department council for record keeping and for permission to assign an F grade in the course. The department policy on academic integrity can be found at: https://minerva.nitc.ac.in/?q=node/650.

Questions

1. You are given a DNA sequence represented as a string S consisting of the characters {A, C, G, T} and a target DNA pattern P. Write a program that uses the KMP algorithm to find all the occurrences of the pattern P in the DNA sequence S. Print the index of the last occurrence and the frequency of P in S.

   Input Format

   - First line contains a string S representing the DNA sequence ($1 \leq |S| \leq 10^6$).
   - The next line contains a string P representing the DNA pattern ($1 \leq |P| \leq 10^5$).

   Output Format

   - Print the index (starts at 0) of the last occurrence of the pattern P in the text S, followed by the number of times the pattern P occurs in S separated by a space. If no match is found, then print -1.

   Sample test case:

   Input:

   ```
   AGCTTAGCTAAGCTT
   AGCTT
   ```

   Output:

   ```
   10 2
   ```

2. You are given a large document represented as a string T and a potentially plagiarized section represented as a pattern P. Implement the Boyer-Moore algorithm to detect whether the pattern P appears in the document T, and if it does, return the first index at which it appears. If the pattern is not found, return `-1`.

   Input Format

   - First part of the input contains a string T representing the large document ($1 \leq |T| \leq 10^7$) and the next part contains a string P representing the pattern ($1 \leq |P| \leq 10^5$) separated by a new line character ('$\backslash n$').

   Output Format

   - Print the index (starts at 0) of the first occurrence of the pattern P in the text T. If no occurrence of the pattern is found, then print `-1`.

   Sample test case:

   Input:

   ```
   Convolutional neural networks (CNNs) are widely regarded as the most effective for image
   classification. They learn spatial hierarchies of features through backpropagation and can
   handle complex visual tasks.

   spatial hierarchies
   ```

   Output:

   ```
   116
   ```