Problem Set 4, Programming Part
String Search

Due: 11:55pm Thursday, November 28, 2017

---

1 Programming Assignment

A geneticist excitedly has just discovered a gene that is a strong predictor for whether or not someone will win the Nobel prize.[1] They immediately wonder whether they themselves have the gene, to help confirm whether their discovery is worthy of the Nobel prize. Therefore, they decide to search for this gene in their own genome, which is provided as a very long DNA sequence of length n. The gene itself is a long sequence of length m. Both sequences take their characters from the alphabet {A, C, G, T}. The geneticist, having little computer science background, would like for you to devise an efficient solution for this problem. All that is required is to identify where the first occurrence of the gene occurs (should it indeed occur), and if it does not occur then to indicate as such. This problem is described by the following Input and Output.

Input:	A text file (the genome) and a pattern (the gene). Both consist entirely of characters from the alphabet {A, C, G, T}.
Output:	If the text contains the pattern, the index of the first occurrence of the pattern in the text. Otherwise, the length of the text.

A Java template has been provided containing one empty function and an empty constructor. The constructor `KMP` takes a string representing the pattern for which we are searching and constructs a DFA. The function `search` takes a string containing the whole text from the given text file and returns the index of the first occurrence of the pattern in the text if the text contains the pattern; otherwise, it returns the length of the text.

You must use the provided Java template as the basis of your submission. You may not change the name, return type, or parameters of those functions. The main function in the template contains code to help you test your implementation by reading it from a file. You may modify the main function or any other function, because your submission will be tested using a different main function. You can use any helper methods or any helper classes. You can use any built-in class or write your own classes and data structures. We advise you to put all the classes you write in the same file, but no other class except the provided one should be declared as a public class.

2 Evaluation Criteria

The programming assignment will be marked out of 40, based on a combination of automated testing (using large texts and patterns) and human inspection.

You are advised to implement the KMP algorithm. For each pattern, the running time of your code should be at most $O(M + N)$, where N is the length of the text being searched and M is the length of the pattern for which you are searching. The mark for your submission will be based on both the asymptotic worst case running time and the ability of the algorithm to handle inputs of different sizes.

[1] The instructor is in no way suggesting that genetics is the sole determinant of intelligence!

Score	Description
0 - 15	Submission does not compile, does not conform to the provided template, or crashes for any of the data sets.
16 - 40	The implemented algorithm is $O(N + M)$ and gives the correct answer on all tested inputs.

To be properly tested, every submission must compile correctly as submitted and must be based on the provided template. If your submission does not compile for any reason (including even trivial mistakes like typos) or was not based on the template, it will receive at most 15 out of 40. The best way to ensure your submission is correct is to download it from conneX after submitting and test it.

You are not permitted to revise your submission after the due date, and late submissions will not be accepted, so you should ensure that you have submitted the correct version of your code before the due date. conneX will allow you to change your submission before the due date if you notice a mistake. After submitting your assignment, conneX will automatically send you a confirmation email. If you do not receive such an email, your submission was not received. If you have problems with the submission process, send an email to the instructor before the due date.