02-714: Homework #1

Due: Sept. 19 at the start of class

Please write your answers neatly or typeset them. You may discuss the problems with your current classmates, but you must write your own solutions entirely independently. If you need to make any assumptions in order to solve a problem, state them explicitly.

1. Let n be the length of a pattern and m be the length of the text. Give an example for each of the following situations:

   (a) Give an example where the version of Boyer-Moore from class takes time $O(nm)$ time.

   (b) Give a pattern of length n and a text of length m where the naïve algorithm takes $O(n + m)$ time.

2. Suppose $Z_2 = q > 0$. What are $Z_3, \ldots, Z_{q+2}$?

3. A string x of length n is a *circular rotation* of y if is can be written as $y[i \ldots n]y[1 \ldots i-1]$. Given two strings x and y, give an algorithm that runs in $O(n)$ time to determine if x is a circular rotation of y.

4. Let x and y be strings of length n and m, respectively. Give an $O(n + m)$-time algorithm to find the longest suffix of x that exactly matches a prefix of y.

5. Without affecting its asymptotic runtime, describe how to support wildcards "?" in the Shift-And algorithm that can match any letter in the pattern or text.

6. Let $A(p, q)$ be the number of characters that match when two strings p and q of length k are compared. Let x and y be strings of length n, and let k be an integer. Each of x and y contain $n - k + 1$ substrings of length k (called k-mers). We want to compute $A(p, q)$ for each pair p, q of kmers where p comes from x and q comes from y. This is easy to do in $O(kn^2)$ time. Give an algorithm to do it in $O(n^2)$ time.

7. Let x be a string of length m and let S be a *multiset* containing n characters. (A multiset is a set where elements can occur more than once.) Give a $O(m)$-time algorithm that finds all occurrences of substrings in x that can be formed using exactly the characters of S. For example, if $S = \{a, a, n\}$ and $x =$ `banana` then the output should be indices 2 and 4.