CIS 667 and CIS 467H: Homework 4(Due: Apr 7, 2005)

Problem 1. (20 points)

(a) Give an input T, P for which NaiveMatching exhibits worst-case performance over a binary alphabet of letters a, b.

(b) Give an input T, P for which Boyer-Moore exhibits worst-case performance over a binary alphabet of letters a, b.

(c) How many comparisons does BM perform if $P = a^m$ and $T = (a^{m-1}b)^{n/m}$.

Assume m/n and a^m means m consecutive occurrences of a.

Problem 2. (10 points)

Let the pattern p be

```
ABRACADABRA
```

and the text t be

```
ABRACAZABRACADABRMABRACADABRA
```

(a) Illustrate the working of the Boyer-Moore algorithm. What is the total number of comparisons (text vs pattern) performed?

(b) Illustrate the working of the Knuth-Morris-Pratt algorithm. What is the f[] array. What is the total number of comparisons (text vs pattern) performed?

(c) Which of the two algorithms is faster for this pattern? Explain.

(d) In Rabin-Karp, let the pattern be $p = 35$ and all arithmetic operations be performed modulo q where $q = 10$. Are there any spurious (false) hits if we search for the pattern p in text t shown below? What is the number of character comparisons (not fingerprint!) performed to find all occurrences of the pattern in the text?

$$2718281828459045235 36$$

Explain.

Problem 3. (20 points)

(a) We want to compute $m = m_1 m_2 \dots m_k$. Show that m can be computed in $O((\lg m)^2)$ bit operations.

(b) Compute $n!$ (n factorial) in $O((n \lg n)^2)$ bit operations.

Problem 4. (20 points)

Let A be an integer. Determine whether A is a perfect power or not in time $O(\lg^2 A)$. An acceptable solution would be one that determines this in slightly more time eg. $O(\lg^3 A)$.

Integer A is a perfect power if there exist integers x, y such that $A = x^y$. **Hint.** $A = x^y$ means that $\lg A = y \lg x$. How big are $y, \lg x$?

Problem 5. (20 points)

(a) Draw the sorting network derived from selection sort and determine its depth. Justify you answer. Assume that selection sort finds in the i-th iteration the maximum of the remaining $n - i + 1$ keys and moves that key in the $n - i + 1$ position in the output sequence.

(b) Draw the sorting network derived from bubble sort and determine its depth. Justify you answer. Bubblesort is on page 38 of CLRS.

Problem 6. (10 points)

You are given a random vector $a = (a_1, \dots a_n)$, where a_i is equally likely and independently to be 0 or 1, i.e. $Pr(a_i = 1) = Pr(a_i = 0) = 1/2$. Answer the following questions.

(a) **(Warmup)** What's the probability that a is the all zero vector (3 points)?

(b) Suppose that a, b are two 0-1 vectors of length n whose components were chosen uniformly at random as discussed previously. What is the expected value of the inner product $a \cdot b = \sum_{i=1}^{n} a_i b_i$? Explain (7 points).

CIS 667 and CIS 467H: Homework 4(Due: Apr 7, 2005)

Option 1. (100 points)

A. Implement the NaiveAlgorithm, RabinKarp, BoyerMoore, and KMP by providing C/C++ implementations **consistent with the descriptions of these algorithms provided in class and in the available notes**. For example implementing descriptions or pseudocode for RabinKarp or any of the other algorithms found on the web will be rejected. String searches should be case insensitive. Then test, your algorithms on texts available at the Project Gutenberg web-site. `http://promo.net/pg/`. Some texts that you should test your algorithms on are

```
1. US Copyright Law at
   ftp://ftp.archive.org/pub/etext/etext03/clusa10.txt

2. George Bernard Shaw's  "Caesar and Cleopatra".
   ftp://ftp.archive.org/pub/etext/etext02/candc10.txt  .
3. Orlando Furioso by Ludovico Ariosto at
   ftp://ftp.archive.org/pub/etext/etext03/8ofur10.txt  .
4. Human Genome Project, Chromosome 1,
   ftp://ftp.archive.org/pub/etext/etext00/01hgp10.txt
```

Search 4-letter, 8-letter and 16-letter strings that appear and do not appear in the text and report the running times for finding all occurrences in a tabular form and the number of successful hits. In addition search the strings "notwithstanding" and "protection" in the first text, "Caesar" and "Cleopatra" in the second, and "Alessandro" and "Alcina" in the third, and "AGCTAGCT" and "AAAAGGGCC" in the fourth, and report the number of hits in addition to running times.