58093 String Processing Algorithms (Autumn 2011)
Exercises 2 (8 November)

1. Outline algorithms that find the most frequent symbol in a given string

   (a) for ordered alphabet, and
   (b) for integer alphabet.

   The algorithms should be as fast as possible. What are their time complexities?

2. Complete the proof of Theorem 1.3 by showing the following result:

   Let $n_1, n_2, \ldots, n_d$ be positive integers, and let $n = \sum_{i=1}^{d} n_i$. Then

$$\sum_{i=1}^{d} n_i \log n_i \geq n \log \frac{n}{d}$$

   Hint: Look up Jensen's inequality.

3. Let R be a multiset containing n elements but only $d < n$ *distinct* elements. Show that ternary quicksort sorts R in $\mathcal{O}(n \log d)$ time. *Hint:* Sum up the maximum number of comparisons for each element and use the result in Problem 2.

4. Let $\mathcal{R}$ be a set of n random strings from Σ^k for some $k > \log_\sigma n$. Show that $dp(\mathcal{R}) = \mathcal{O}(n \log_\sigma n)$ on average.

5. Let $\mathcal{R} = \{S_1, S_2, \ldots, S_n\}$ be a (multi)set of strings such that $S_1 \leq S_2 \leq \cdots \leq S_n$. Define the LCP array $LCP_{\mathcal{R}}[2..n]$ as $LCP_{\mathcal{R}}[i] = lcp(S_{i-1}, S_i)$. Let $lcp(\mathcal{R}) = \sum_{i=2}^{n} LCP_{\mathcal{R}}[i]$. Show that

$$lcp(\mathcal{R}) \leq dp(\mathcal{R}) \leq 2 \cdot lcp(\mathcal{R}) + n \; .$$

6. An integer can be seen as a string of digits; the standard decimal notation is an example. On the other hand, a string over an integer alphabet can be interpreted as an integer expressed in base-σ notation. Let $I(S)$ be the value of this integer for a string S, i.e., $I(S) = \sum_{i=0}^{|S|-1} S[i] \cdot \sigma^{|S|-i-1}$.

   (a) This interpretation induces an order on strings, namely the order $A \preceq B$ if and only if $I(A) \leq I(B)$. Give a definition of this order in terms of strings without referring to the integer interpretation (i.e., something similar to the definition of the lexicographical order in the lecture notes).
   (b) A string S can also be interpreted as the rational number $I(S)/\sigma^{|S|} \in [0, 1)$. Is the corresponding induced order the same as the lexicographical order?