58093 String Processing Algorithms (Autumn 2010)
Exercises 3 (25 November)

1. A q-gram of a string is its factor of length q. For example, the 3-grams of the string `ararat` are `ara`, `rar`, `ara` and `rat`. Show that if $ed(A, B) \leq k$, then the strings A and B have at least $|A| - q + 1 - kq$ common q-grams.

2. Outline a filtering algorithm based on the result in Problem 1.

3. Complete the proof of Theorem 3.2 by showing the following result:

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

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

   Hint: Look up Jensen's inequality.

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

5. As mentioned on the lectures, an integer can be seen as a string of digits. 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.

   (a) This interpretation induces an order on strings, namely the order $A \lesssim B$ 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 this the lexicographical order?

6. Describe how to modify the LSD radix sort algorithm to handle strings of varying lengths. The time complexity should be the one given in Theorem 3.13.

7. Use the lcp comparison technique to modify the standard insertion sort algorithm so that it sorts strings in $\mathcal{O}(DP(\mathcal{R}) + n^2)$ time.