Data Structures - Assignment no. 9, June 27, 2007

Remarks:

- **Write both your name and your ID number very clearly on the top of the exercise. Write your exercises in pen, or in clearly visible pencil. Please write *very* clearly.**
- **Recall that 80% of the theoretical exercises must be submitted. The exercises can and must be worked on and submitted alone.**

1. Build a Huffman tree T for a string s_1 which contains two 'A's, three 'B's, eight 'C's, ten 'D's, and twelve 'E's. Then compress the string $s_2 =$"ABDE" using the encoding represented by T. What is the size, in bits, of this encoding of s_2. What is the size, in bits, of the encoding of s_1?

2. Draw a suffix tree for the string "mississippi". Then show the search path for the sub-string "ssi" and for the sub-string "ssip".

3. Let s be a string of length 10^6, which consists of $0.99 \cdot 10^6$ 'A's and $0.01 \cdot 10^6$ 'B's.

   (a) Build a Huffman tree for this string.

   (b) Calculate the length, in bits, of the Huffman encoding of the string.

   (c) Can you come up with a better method of encoding the string by a sequence of bits? Hint: $\log_2(10^6) \approx 20$.

   (d) (Optional. This will not be checked, but we encourage you to think about it anyway) Why didn't Huffman encoding give good results here? What drawback of Huffman encoding can you point out? Can you think of a way to avoid this problem?

4. (Kraft's Inequality)

   (a) Let T be a binary tree. For a leaf v of T, let $depth(v)$ be the depth of v, measured in edges. Prove that for any such tree T, $\sum_v 2^{-depth(v)} \leq 1$, where the sum is taken over all leaves v. (Hint: Use induction on the number of nodes in the tree).

   (b) Suppose we work over alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$. Let PF be a prefix-free code that encodes σ_i into ℓ_i bits. Prove that for any such prefix-free code PF, $\sum_{i=1}^{k} 2^{-\ell_i} \leq 1$. (Hint: this is a direct consequence of (a)).

   (c) Suppose we work over alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$. Let $\ell_1, \ldots, \ell_k$ be positive integers such that $\sum_{i=1}^{k} 2^{-\ell_i} \leq 1$. Prove that there exists a prefix-free code PF that encodes σ_i into ℓ_i bits.
   Hint: First, note that you can assume without loss of generality that $\sum_{i=1}^{k} 2^{-\ell_i} = 1$. Now, prove the claim by induction on k. You would probably want to use the fact that if ℓ_i is the maximal element among $\ell_1, \ldots, \ell_k$, then it is easy to see that there is $j \neq i$ such that $\ell_j = \ell_i$. This follows from the fact that the binary encoding of $2^{-\ell_i}$ has '1' in its ℓ_i-th bit, and in order for $\sum_{i=1}^{k} 2^{-\ell_i} = 1$ to hold, there must be another ℓ_j with '1' in its ℓ_i-th bit.