CS 38 Introduction to Algorithms **Spring 2014**

Problem Set 2

Out: April 15 *Due:* **April 22**

Reminder: you are encouraged to work in groups of two or three; however you must turn in your own write-up and note with whom you worked. You may consult the course notes and the optional text (CLRS). The full honor code guidelines can be found in the course syllabus.

Please attempt all problems. **To facilitate grading, please turn in each problem on a separate sheet of paper and put your name on each sheet. Do not staple the separate sheets.**

1. Recall that a prefix-free encoding scheme can be represented by a binary tree, and that the Huffman code algorithm gives an efficient way to construct an optimal such tree from the probabilities $p_1, p_2, \ldots, p_n$ of n symbols. In this problem, you will show that the average length of such an encoding scheme is at most one larger than the *entropy* (which is the information-theoretic best-possible). The *entropy* of the distribution given by $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ is defined to be

$$H(\mathbf{p}) = \sum_{i=1}^{n} -\log(p_i) \cdot p_i.$$

(a) Prove that for any list of positive integers $\ell_1, \ell_2, \ldots, \ell_n$ satisfying

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1,$$

there is a binary tree with distinct root-leaf paths having lengths $\ell_1, \ell_2, \ldots, \ell_n$. Hint: start from the full binary tree and delete subtrees.

(b) Let $\mathbf{p} = (p_1, \ldots, p_n)$ give the probabilities of n symbols. Prove that if $\ell_1, \ldots, \ell_n$ are the encoding lengths in an optimal prefix-free encoding scheme for this distribution, then the average encoding length,

$$\sum_{i=1}^{n} \ell_i p_i,$$

is at most $H(\mathbf{p}) + 1$.

2. A *matroid* is a family $\mathcal{I}$ of subsets of a universe E, satisfying the following three axioms:

- $\mathcal{I}$ contains the empty set,
- if $A \in \mathcal{I}$ and then every subset $B \subseteq A$ is also in $\mathcal{I}$, and
- if A and B are subsets in $\mathcal{I}$ with $|B| > |A|$, then there exists $x \in B \setminus A$ such that $A \cup \{x\}$ is in $\mathcal{I}$.

The subsets in $\mathcal{I}$ are called *independent sets*. The maximal independent sets are called *bases*. A consequence of the third axiom is that all of the bases have the same cardinality.

(a) Two standard examples of matroids are the *graphic matroid* which is defined relative to a graph G, and whose independent sets are all subsets of the edges of G that are forests, and the *matric matroid* which is defined relative to an $n \times m$ real matrix M, and whose independent sets are all subsets of columns of M that are linearly independent. Prove that these two structures are indeed matroids.

(b) Given a matroid $\mathcal{I}$ and nonnegative integer weights for each element of the universe E, we want to find an independent set of *maximum* weight (the weight of a subset is the sum of the weights of its elements). Give a greedy algorithm for this problem that runs in time $O(|E| \log |E|)$ plus $O(|E|)$ calls to a procedure that determines whether or not a given set A is independent, and prove that it is correct.

(c) The *dual* of a matroid $\mathcal{I}$ over a universe E is the family of subsets $\mathcal{I}'$ over the same universe E, consisting of the complements of bases of $\mathcal{I}$, and all subsets of these sets. Prove that the dual $\mathcal{I}'$ is itself a matroid.

(d) Consider a matroid $\mathcal{I}$ over a universe E, and for a parameter k, define the following family of subsets

$$\mathcal{I}_k = \{A : A \in \mathcal{I} \text{ and } |A| \leq k\}.$$

Prove that for each positive integer k, $\mathcal{I}_k$ is a matroid.

3. Here are some problems that can be solved greedily, via formulating them as matroids.

(a) Formulate the minimum cost spanning tree problem as the problem of finding a maximum weight independent set in a matroid.

(b) A *pseudo-forest* in an undirected graph is a subset of the edges containing at most one cycle. Formulate the problem of finding a maximum weight pseudo-forest in an undirected graph with non-negative edge weights as the problem of finding a maximum weight independent set in a matroid.

(c) An Internet Service Provider maintains a network of links described by a connected undirected graph. Each link has a yearly cost associated with it. The company decides to decommission k links, but it needs the graph to remain connected. Formulate the problem of finding which k links to retire so as to maximize the yearly savings as the problem of finding a maximum weight independent set in a matroid. Hint: you may want to use the last two parts of the previous problem.

4. We are given a rooted tree T on n nodes and m pairs of vertices. For each pair, we wish to output its least common ancestor in the tree T. Give an algorithm solving this problem, that runs in time $O((n+m)\log^* n)$, using the Union-Find data structure and the "$\log^* n$ analysis" from class. Hint: perform a DFS of the tree, union-ing subtrees as you encounter them until the entire tree constitutes one set in the union-find data structure.