CSE1729	Introduction to Programming Problem Set 10	April 15, 2017

Constructing Huffman trees

Recall the discussion from class on Huffman trees. In particular, to construct an optimal encoding tree for a family of symbols $\sigma_1, \ldots, \sigma_k$ with frequencies $f_1, \ldots, f_k$, carry out the following algorithm:

1. Place each symbol σ_i into its own tree; define the weight of this tree to be f_i.
2. If all symbols are in a single tree, return this tree (which yields the optimal code).
3. Otherwise, find the two current trees of minimal weight (breaking ties arbitrarily) and combine them into a new tree by introducing a new root node, and assigning the two light trees to be its subtrees. The weight of this new tree is defined to be the sum of the weights of the subtrees. Return to step 2.

As an example, consider Huffman encoding a long English text document:

- You would begin by computing the frequencies of each symbol in the document. This would produce a table, something like the one shown below.

Symbol	Frequency
a	2013
b	507
c	711
⋮	⋮

  Here the "frequency" is the number of times the symbol appeared in the document. (If you prefer, you could divide each of these numbers by the total length of the document; in that case, you could think of the frequencies as probabilities. This won't change the results of the Huffman code algorithm.)

- Following this, you can apply the Huffman code algorithm above: this will produce a "Huffman code tree." The purpose of this tree is to associate a "codeword" with each symbol. Specifically, the path from the root to a given symbol can be turned into a codeword by treating every step to a left child as a zero and every step to a right child as a one. In the figure below, the symbol α would be associated with the codeword 100.

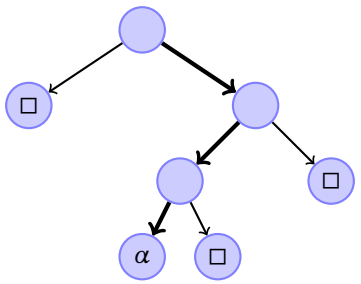

Figure 1: A Huffman tree; the codeword associated with α is 100. The □ placeholders represent other symbols.

- Finally, you can "encode" the document by associating each symbol in the document with the corresponding codeword. Note that this will turn the document into a long string of 0s and 1s.
- Likewise, you can "decode" the encoded version of a document, by reading the encoded version of the document from left to right, and following the path it describes in the Huffman tree. Every time a symbol is reached, the process starts again at the root of the tree.

Strings and characters in SCHEME SCHEME has the facility to work with strings and characters (a *string* is just a sequence of characters). In particular, SCHEME treats *characters* as atomic objects that evaluate to themselves. They are denoted: `#\a`, `#\b`, ... Thus, for example,

```
> #\a
#\a
> #\A
#\A
> (eq? #\A #\a)
#f
> (eq? #\a #\a)
#t
```

The "space" character is denoted `#\space`. A "newline" (or carriage return) is denoted `#\newline`.

A *string* in SCHEME is a sequence of characters, but the exact relationship between strings and characters requires an explanation. A string is denoted, for example, `"Hello!"`. You can build a string from characters by using the `string` command as shown below. An alternate method is to use the `list->string` command, which constructs a string from a list of characters, also modeled below. Likewise, you can "explode" a string into a list of characters by the command `string->list`:

```
> (string #\S #\c #\h #\e #\m #\e)
"Scheme"
> (list->string '(#\S #\c #\h #\e #\m #\e))
"Scheme"
> (string->list "Scheme")
(#\S #\c #\h #\e #\m #\e)
> "Scheme"
"Scheme"
```

Note that strings, like characters, numbers, and Boolean values, evaluate to themselves.

---

1. Write a SCHEME function which, given a list of characters and frequencies, constructs a Huffman encoding tree. You may assume that the characters and their frequencies are given in a list of pairs: for example, the list

   ```
   ((#\a . 2013) (#\b . 507) (#\c . 711))
   ```

   represents the 3 characters *a*, *b*, and *c*, with frequencies 2013, 507, and 711, respectively. Given such a list, you wish to compute the tree that results from the above algorithm. I suggest that you maintain nodes of the tree as lists: internal nodes can have the form

   ```
   (internal 0-tree 1-tree)
   ```

   where `internal` is a token that indicates that this is an internal node, and `0-tree` and `1-tree` are the two subtrees; leaf nodes can have the form

   ```
   (s () ())
   ```

where `s` is the character held by the leaf. Note that you will have to use the SCHEME `quote` command to construct internal nodes: for example, to construct an internal node with the two subtrees `0-tree` and `1-tree`, you could use the procedure below

```
(define (make-internal-node 0-tree 1-tree)
  (list 'internal 0-tree 1-tree))
```

Note that when you traverse a Huffman coding tree, you can determine if a given node is an internal node by deciding if the `car` of the list associated with that node is the token `internal`.

2. Define a SCHEME function that takes, as input, a Huffman coding tree and outputs a list containing the elements at the leaves of the tree along with their associated encodings as a string over the characters `#\0` and `#\1`. For example, given the tree of Figure 2, your function should return the list

```
((#\a . "0") (#\b . "10") (#\c . "11")) .
```

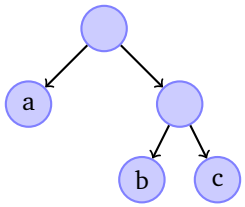

Figure 2: A Huffman tree yielding an encoding of the three symbols a, b, and c.

3. Define a SCHEME function `encode` which takes, as input, a string and a list of characters and frequencies (as in problem 1) and encodes the string into a 0/1-string using Huffman coding. (You may assume that every character in the string is indeed in the list of character/frequency pairs.)

4. Define a SCHEME function `decode` which takes, as input, a string over the symbols 0/1 and a Huffman coding tree and "decodes" the string according to the tree. (It should return a string.)

5. Finally, show how to package all of the functionality you have created in this assignment as a `Huffman-code` object. Here's the idea.

   - Define a function `H-code-object` which takes, as an argument, a list of character/frequency pairs. It should return a dispatcher that yields two methods: `encode`, which maps a character string to a string of zeros and ones, and `decode`, which maps a string of zeros and ones to a character string. You might also implement a "reminder" method which simply returns the set of allowable characters (the ones that appeared in the original character/frequency list).
   - Your method should maintain a Huffman encoding