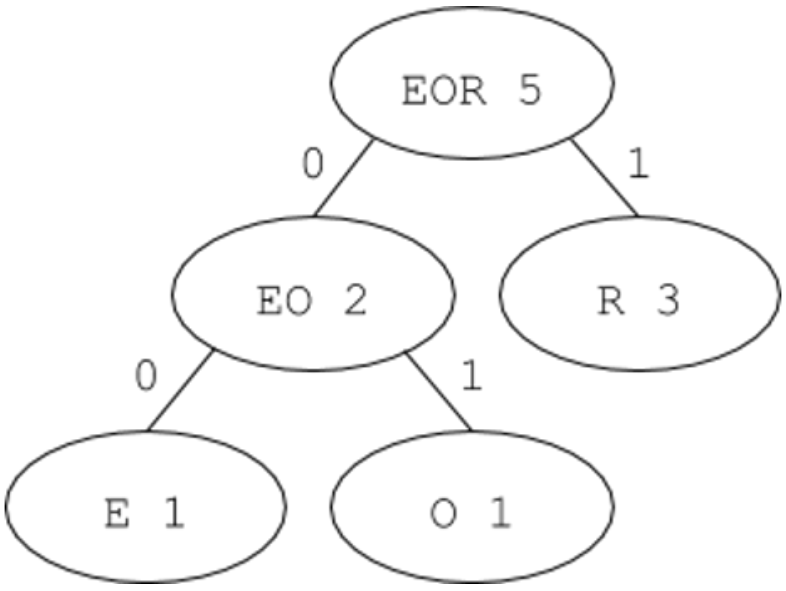
Huffman Encoder Part 1: Building a Huffman Tree

In this assignment, we will implement an efficient scheme for compressing a text message called Huffman coding. A simple method to encode text as a string of 0s and 1s is to represent each character as a unique string of binary digits (bits). A text message is thus translated into a string of bits. By exploiting the fact that not all characters appear with the same frequency in the text, we can encode rarely used characters with long codes and frequently used ones with short codes.

Here, we'll use a given a set of characters and their corresponding frequencies to create an optimal coding scheme: a special binary tree where the path to each leaf represents a different character.

Huffman Tree

A Huffman Tree stores elements based on their frequency such that the higher frequency characters have shorter paths. For example, let's take the word "ERROR". The character/frequency legend is `E 1 R 3 O 1` and here's a corresponding Huffman Tree:



The leaf nodes are the characters with their frequencies, and the internal nodes are combinations of its children's characters with the sum of their frequencies. You should also notice the labels on the paths from node to node (each left child is a 0 and each right child is a 1); these are the basis for encoding.

If you follow the path from the root to each leaf, concatenating the edge labels, you get a unique string for each character. `E` is encoded as `00`, `O` is `01`, and `R` is simply `1`. The more frequent character `R` has a shorter string and so we can encode the word "ERROR" with only 7 bits (the encoding is `0011011`). This is far fewer than ASCII.

Implementation

We'll use a binary heap to generate a Huffman Tree from the character frequencies. Set up a project/package with the following classes:

```
public class BinaryHeap // the heap class as posted
{
  // public int getSize()
}

public class HuffmanNode implements Comparable<HuffmanNode>
{
   public String letter;
   public Double frequency;
   public HuffmanNode left, right;

   // public HuffmanNode(String letter, Double frequency)
   // public HuffmanNode(HuffmanNode left, HuffmanNode right)
   // public int compareTo(HuffmanNode o)
   // public String toString()
}

public class HuffmanTree
{
   HuffmanNode root;

   // public HuffmanTree(HuffmanNode huff)
   // public void printLegend()
   // private void printLegend(HuffmanNode node, String code)
   // public static BinaryHeap<HuffmanNode> legendToHeap(String legend)
   // public static HuffmanTree createFromHeap(BinaryHeap<HuffmanNode>
   heap)
   // public static void main(String[] args)
}
```

Here's information about the above methods:

BinaryHeap

`public int getSize()`. This method returns the number of elements in the heap.

HuffmanNode

`public HuffmanNode(String letter, Double frequency)`. This constructor creates a new HuffmanNode where letter is set to *this.letter*, frequency is set to *this.frequency*, and *left* and *right* are set to null.

`public HuffmanNode(HuffmanNode left, HuffmanNode right)`. This constructor creates a new HuffmanNode from its two children (i.e. the two nodes passed as parameters should become children of the new node), setting the *letter* variable to the concatenation of *left.letter* & *right.letter*, and the frequency variable to the sum of *left.frequency* & *right.frequency*.

`public int compareTo(HuffmanNode huff)`. This return `this.frequency.compareTo(huff.frequency)`. This allows us to make a heap of HuffmanNodes where the frequency determines which node is larger than which.

`public String toString()`. It returns a string of form `"<"+letter+", "+frequency+">"`. There's no need to recursively iterate left/right pointers in this method.

HuffmanTree

`public HuffmanTree(HuffmanNode huff)`. This constructor sets *this.root* to *huff*.

`public void printLegend()`. This calls `printLegend(root, "")`, which calls private `void printLegend(HuffmanNode node, String code)`, a recursive method that works as follows: If `(node.letter.length() > 1)` i.e., *node* contains multiple characters, then *node* is NOT a leaf node, so we recursively call printLegend() on its left child using printLegend(`node.left`, `code + "0"`), and recurse on *node*'s right child using printLegend(`node.right`, `code + "1"`). If *node.letter* is a single character, then *node* is a leaf node, and we print out `(node.letter+"="+code);`

`public static BinaryHeap legendToHeap(String legend)` Converts a string legend into a binary heap of Huffman nodes. The legend string contains pairs of characters and their frequencies, separated by spaces. For example, "A 20 E 24 G 3" means that character * 'A' has a frequency of 20, 'E' has a frequency of 24, and 'G' has a * frequency of 3. This method splits the string into parts, then creates * a Huffman node for each pair and inserts it into a binary heap.

`public static HuffmanTree createFromHeap(BinaryHeap b).` We run the Huffman algorithm here. When we have only one element left in the heap, we remove it, and create a new HuffmanTree object with *root* set to the removed object.

`public static void main(String[] args)` calls `legendToHeap()` on the legend string and returns a BinaryHeap. We then call `heap.printHeap()` on the heap. Next, we call `createFromHeap(heap)` on the heap to run our Huffman algorithm which returns a HuffmanTree, called, here, *huffmantree*. Finally, we call `huffmantree.printLegend()` on this HuffmanTree object to print the binary encodings for each of the letters in our input file.

The Algorithm

The input is a legend of characters and their corresponding frequencies. The output is a Huffman Tree, built using a Binary Heap.

1. Create a single HuffmanNode for each letter and its frequency, and insert each of these into a new BinaryHeap.
2. While the Binary Heap has more than one element:
    a. Remove the two nodes with minimum frequency.
    b. Create a new HuffmanNode with those minimum frequency nodes as children (using the the HuffmanNode constructor with left and right nodes as parameters) and insert that node into the BinaryHeap.
3. The BinaryHeap's only element will be the root of the Huffman Tree. Pass this node into the HuffmanTree constructor and return the result.

Legend

The test data for part 1 of this program is:

```
A 20 E 24 G 3 H 4 I 17 L 6 N 5 O 10 S 8 V 1 W 2
```

```
<V, 1.0> <W, 2.0> <N, 5.0> <S, 8.0> <G, 3.0> <A, 20.0> <L,
6.0> <E, 24.0> <O, 10.0> <I, 17.0> <H, 4.0>
A=00
O=010
G=01100
V=011010
W=011011
L=0111
E=10
I=110
S=1110
H=11110
N=11111
```

It's ok to hardcode this string. You can and should try out your program with other character/frequency legends.

Submission

You will submit your HuffmanNode, HuffmanTree, and adjusted BinaryHeap classes with the detailed methods. You should also submit the unchanged UnderflowException file.