CMPSCI 145
Huffman Compression
Professor William T. Verts

This is the last computer assignment for CMPSCI 145. You may find this assignment to be a bit tedious in spots, so please get started on it early.

1. The following text is the first paragraph of Abraham Lincoln's Gettysburg Address. The left column of the text is column 1 (i.e., there are no characters to the left of the text that you see). The **??** characters are <u>not question marks</u>, but are used here to denote the carriage-return (ASCII code 13) followed by the line-feed (ASCII code 10) at the right end of each line, characteristic of text files on Windows PCs. Carefully count each type of character (how many **A**s, how many **B**s, etc.) and write that information down. Don't forget to count the carriage-returns, line-feeds, commas, periods, and spaces. The total is exactly 184 characters.

```
FOUR SCORE AND SEVEN YEARS??
AGO OUR FATHERS BROUGHT??
FORTH ON THIS CONTINENT,??
A NEW NATION, CONCEIVED IN??
LIBERTY, AND DEDICATED TO??
THE PROPOSITION THAT ALL??
MEN ARE CREATED EQUAL.??
```

2. Sort your frequency counts in descending order by letter frequency. Use the Huffman code technique shown in class to generate, ***by hand***, a single Huffman tree of the characters. You'll have to do this manually. Be very careful!

3. Once you have your final tree design drawn out on paper, download from the class web site the latest version of a program called "Build Huffman" and unpack the **BuildHuffman_Distribution.ZIP** file into a folder on your hard disk (the archive contains a single **.EXE** file called **BuildHuffman.exe**).

4. Launch the program and then maximize the window. Play with the program for a while to get a feel for how it works. Click on both round (internal) and square (leaf) nodes, type in values, and use the arrow buttons in the tool panel to generate new sections of the tree. When you are done, click on the root node and hit Del enough times to delete the entire structure down to the original configuration.

5. Use the BuildHuffman program to create an identical copy of the tree you drew out on paper. For each square (leaf) node, type in the appropriate character, then hit F2 (or click the % button) and type in the corresponding frequency count. For carriage-returns use the abbreviation **CR**, for line-feeds use **LF**, and for spaces use **SP**. Each round (internal) node will show the total of all nodes below it in the tree; when the tree is complete, the root (top) node will show the total number of characters in the file. Verify that this number matches the total number of characters counted in step 1 (184).

6. Save the text (File-Save Text...) to a text file with your last name, an underscore, the initial letter of your first name, and the string **_HUFFMAN** as the name of the file (I would use **VERTS_W_HUFFMAN.TXT**). Similarly, save the image (File-Save Image...) to a bitmap file with the same pattern and a **.BMP** extension as the name of the file (I would use **VERTS_W_HUFFMAN.BMP**).

7. Compute the total number of bits in the original file by multiplying the total number of characters by 8 (the number of bits in an ASCII character). Look at the Total Bits entry in the text panel of **BuildHuffman**, and then compute the compression factor by dividing the uncompressed total by the compressed total. Using Windows Notepad, edit your **.TXT** file created in step 6 by opening up some space at the bottom of the file and typing in the compression factor you just computed. Save the edited file.

8. Package your **.TXT** and your **.BMP** files into a new **.ZIP** archive with the string **HUFFMAN**, an underscore, your last name, an underscore, and the initial letter of your first name as the name of the archive (for example, I would create an archive called **HUFFMAN_VERTS_W.ZIP**). Email the **.ZIP** archive to the standard mailbox at: **literacy@cs.umass.edu** as an email attachment. In your email message set the subject line to:

CMPSCI 145 Huffman