Assignment #8 (Implementation of a compression software with Huffman algorithm)

Goal: Use the Huffman algorithm to develop a software tool for data compression and decompression.

Guide:

Data compression is an approach to reduce the required amount of storage or transmission, which has extensive applications. The techniques of data compression can be categorized into two main kinds, lossy compression and lossless compression. The lossless compression, in which the original data can be retrieved exactly by decompressing the compressed data, is usually adopted for file compression. The lossy compression, in which the decompressed data may be slightly different from the original one, is usually used for multi-media compression.

The Huffman compression is a lossless algorithm, whose source code can be found in the text book. In this assignment, you can use the program of Huffman and the function of min heap from the textbook. It is not viewed as plagiarism.

Steps for data compression:

Step 1：For the file to be compressed, compute the frequency of each character, where each character is stored in a byte.

Step 2：Construct the encoding table with Huffman algorithm.

Step 3：Store the encoding table in the header of the compressed file

Step 4：Use the encoding table to encode (compress) each character in the original file, and store the encoded data in the compressed file (following the header).

Step 5：Compute the compression ratio (uncompressed size / compressed size). In the header of the compressed file, please store the size of the original file, the size of the compressed file (including the header), and the compression ratio.

Specification for your Huffman code:

To regulate the produced Huffman code, you have to obey the following rules:

Rule 1: In your Huffman tree, the left subtree is labeled with 0, and the right subtree is labeled with 1.

Rule 2: When two nodes are merged, the node of less lexical order is set as the left son, and the other is set as the right son (the lexical order is decided by the least symbols in two nodes).

Rule 3: If there are more than two nodes of the same least frequency, then you have to merge the two with the least and the second least symbol in lexical order.

To decompress a compressed file, you can refer to the encoding table in the header to retrieve each character. This assignment is operated in text mode, and should have two operations: compress file and uncompress file. Besides, your program must provide correct functions: open file, read file and write file, and you have to output your Huffman encoding table to a txt file.

Example of operation:
huffman –c –i infile –o outfile
(compress operation , infile is input file, and outfile is output file)
huffman –u –i infile –o outfile
(uncompress operation, infile is input file, and outfile is output file)

Notice:

- When executing the program, you need to print the header data into standard output. The output must contain the amounts of the bytes in the original file, the amounts of the bytes in the compressed file, compress ratio and the encoding table.
- In the assignment, we will use the command line to execute. You can read the input from setting parameters in main(int argc, char *argv[]).
- To handle general files (may not be simple texts, such as jpg or mov), please open the file with the binary mode.

Sample text files (The left part is the input, and the right part is the output for the Huffman encoding table

ABCABCADC	A=00 B=010 C=1 D=011
DGHFDFGCJBCHAJDGAGCGBI HJCACDHFJCCIJHDGBCJDBJE HIEFADGBAGDCAGBFFEHGF GJDHEFCFEHEEDAHIGIIEDBE JCAGIJCGGEJFAEIHBHCHCCG BAGBFJEJAJBHHBBFJHHGHD GFGDJEIBCJDJAIIFAJGDIFGBB FJIEIBADAEEEEHGDBHFJGCB HIFHEBIFJHFDIIDBIJGEDDDIG AEIIIJHHGEDADAJCFEAFJEGJ CCEHHFEIBCIJDGIACBAFBBA IFAGEGCFIGCIEHCEAGCJJAIB DBBADEADJBJGJDDIAFGGEA AEGCGHFGJIHCICJJGACHGBD CBFJJBEJBAHFIHAEIAECJGBA ECCCAFFCJACIDIAEAJBJGCFE EJJDGAFACJEDFBHEDBEGJEC CGEBCIIJCIIIGIDIDDDIAEHGIJ AGIHJADDDHDCGFCGFIHJGA EBAEIHAIGCEHJDJCIDABHIJB FEJCFHEJJEFJDFHAIBJJHICFD EJACABJGCBCAHDBBJAFGH	A=000 B=1000 C=010 D=1010 E=011 F=1001 G=110 H=1011 I=001 J=111

About the assignment:

1. Submit the assignment through the website.
2. You will need to demo in EC0513 personally
3. In addition to the above sample, TA will test your program with his own testing data.
4. You will get bonus score if you implement Graphical User Interface (GUI).