CSE3081: Design and Analysis of Algorithms (Fall 2023)
Machine Problem 3: Huffman Coding for File Compression

Handed out: November 21, Due: December 11, 11:59PM (KST)

1. Goal

The goal of this MP is to design and implement a Huffman Coding-based file compression utility program.

2. Problem Description

File compression is important when you want to save disk space or send files through the network. You want to write a utility program which you can use to compress a file and recover the original file from a compressed file.

For compression, you are going to use Huffman Coding, which is a greedy algorithm. The goal of the algorithm is to assign long bit strings to rare characters and short bit strings to characters that appear frequently. Refer to the lecture slides for details.

3. Your task and requirements (Read Carefully!)

(1) You will write a single C/C++ program which takes an input file and produces an output file. Your file should have two functions: compress and decompress. The user command indicates whether we are compressing (-c) or decompressing (-d).

(2) When you compress, your output file will have ".zz" appended at the end of the input file name. If the user runs the program as follows:

```
$ mp3_20210001 -c input.txt
```

Then, your output file should be "input.txt.zz".

(3) When you decompress, your output file will have ".yy" appended at the end of the input file name. If the user runs the program as follows:

```
$ mp3_20210001 -d input.txt.zz
```

Then, your output file should be "input.txt.zz.yy".

(4) If the user gives options other than "-c" or "-d", the program should print an error message and stop. Also, if the input file does not exist, the program should print an error message and stop.

(5) In this MP, you do not need to worry about whether the user will try to decompress a non-compressed file. The TA will compress the file using your program and decompress the output of the compression using your program.

For example, the TA will run commands in (2) and (3), and compare "input.txt" and "input.txt.zz.yy". Obviously, they should be exactly the same.

(6) When you decompress, you should NOT use any other input file except the compressed file itself. In other words, you should not create a separate file which will help you in decompression.

(7) The input files will be text files. Specifically, you can safely assume that all characters used in the file are from the ASCII code table. (http://www.asciitable.com/) No Korean (or non-English characters) will be used. However, we will test your program with files that have different distribution of characters (e.g. A file that mostly consists of numbers, and a file that mostly consists of alphabets.)

(8) The format of the compressed file is up to you. You may embed any additional information necessary for decompression in the compressed file. However, watch out for the size of the compressed file. That is the performance of your algorithm.

(9) When evaluating your work, we are more interested in the size of the compressed file. For performance in terms of time, you will not be deducted points unless your program takes unreasonably large amount of time for compressing or decompressing a file.

(10) Use Huffman Coding! Do not use other compression algorithms you find on the Internet.

(11) Similar to mp1 and mp2, your code should build and run on a **Linux machine**. So make sure you test on the machine before you submit the files.

(12) You should write a **Makefile** this time too. The TA will build your code by running 'make'. It should create the necessary binary file.

(13) Your binary file should be named **mp3_20210001**. The red part should be your student ID. There should be only a single binary file. It is up to you to make a single or multiple source code files.

4. Submission

You should submit the Makefile and the source code(s). Make the file into a zip file named cse3081_mp3_**20210001**.zip. The red numbers should be your student ID. You can submit your work on the cyber campus.

5. Evaluation Criteria

(1) Correctness of your implementation: 70%

- Does your implementation produce correct results? When we compress a file and then decompress the compressed file, do we get the original file back?

※ The maximum size of the input file will not exceed 100 megabytes.

(2) Performance of your program: 30%

If your program produces wrong result, you will get 0 points here. Otherwise, the points will be given based on the size of the compressed file (after comparing performance with other students.) In addition, if your program takes unreasonably long time, that will also be counted towards the performance evaluation.

※ The input files will include multiple files with different sizes.

6. Notes

- You should write your own code. You can discuss ideas with other students, but definitely should not copy their work. For this particular MP, since you can see the example code on the lecture slides, you will lean towards following the code exactly. My advice is that you just catch the idea, and start writing your own code without looking at the example code.

- As announced in the first class, duplicates will receive zero grade.

- You may write your program in your own environment (Windows, Linux, MAC). But you should test your code on the Linux machine before submitting the file. Each of you will be given an account to the department Linux server.

- Remember that the TA will place your files in a directory, build your code using ‘make’, and run the code with the test inputs. Make sure everything works before submitting your work.

- Do NOT submit binary files. Submit only the files listed in the Submission section.

7. Late Policy

- 10% of the score is deducted for each day, up to three days. Submissions are accepted up to three days after the deadline.