Topics in Data Science, Data Protection — Spring 2020
CS 4364 — L. Longpré

Assignment 2, due 3/13/2020

In this assignment, you will write a program that uncompresses a file that was compressed with the algorithm seen in class. An updated Java program that compresses a file is provided. You may write the uncompress program in the language of your choice, as long as it uncompresses the file back to its original form.
The program that compresses the file works as follows:

1. In an array of size 257, read the input file and compute the frequency of occurrence of each byte. The 257th position is used for an end-of-file "character" and its frequency is set to 1.

2. Compute a Huffman code (as a binary tree) based on the frequency of the bytes (and eof) that occur at least once in the input file.

3. Create an array of size 257 containing the Huffman code of each occurring byte.

4. Generate the output file as follows:

   (a) Output a byte with the total number $n$ of different bytes occurring in the input file, counting the eof character.

   (b) Output each byte occurring in the input file, in the order it appears in a traversal of the Huffman code tree. The eof symbol is represented by two consecutive zero bytes, so this actually generates $n + 1$ bytes.

   (c) Output the tree structure as a sequence of bits packed in bytes, padding the last byte with zeroes if the number of bits is not a multiple of 8.

   (d) Output the concatenation of the Huffman codes of all characters from the input file packed in bytes, ending with the Huffman code of end-of-file, padding the last byte with zeroes if the number of bits is not a multiple of 8.

We have posted the updated *Node.java* and *Main.java*. *Heap.java* has not changed.
In *Node.java* only the format for printing was updated from `0x%02X` to `0x%03X` to account for eof having 9 bits.
*Main.java* had the following modifications:

- The end-of-file symbol is represented by 0x100, which does not fit in a byte. Because it does not fit in a byte, the symbol is represented as two consecutive zero bytes in the compressed file.
- Arrays `freq, posfreq` and `values` incremented their size to account for the eof symbol
- Some loops upper bound were also incremented
- Method `outputBytes` has a special case for eof (output two zero bytes.)
- Some values printed with format `0x%03X` instead of `0x%02X`.
- The inadequate way of encoding the end of file was removed, and the end-of-file code is now added at the end of `outputContents()`.

## Submission

Submit by sending an e-mail to *longpre@utep.edu* with subject line *CS4364 Assignment 2* containing the following:

1. Your program
2. The uncompressed version of the compressed file that was sent to you by e-mail.
3. The answer to the following questions:
   (a) How long did you spend on this assignment
   (b) Any difficulty you encountered while doing this assignment, and how you resolved your difficulty
   (c) Any reference you consulted for this assignment

## Early/Late submission

9% bonus points if submitted by due date Friday 3/13 11:59pm.

1% point penalty for each day late if submitted by Monday 3/23 11:59pm.