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

Assignment 2, due 3/8/2020

In this assignment, you will write a program that uncompresses a file that was compressed with the algorithm seen in class. A 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 256, read the input file and compute the frequency of occurrence of each byte.
2. Compute a Huffman code (as a binary tree) based on the frequency of the bytes that occur at least once in the input file.
3. Create an array of size 256 containing the Huffman code of each occurring byte.
4. Generate the output file as follows:
   (a) Output a byte with the number of bytes occurring in the input file.
   (b) Output each byte occurring in the input file, in the order it appears in a traversal of the Huffman code tree.
   (c) Output the tree structure as a sequence of bits packed in bytes, padding 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, padding the last byte with zeroes if the number of bits is not a multiple of 8.
   (e) Possibly output 0-valued bytes to make sure the uncompress program does not confuse the padding with the code for an extra byte.