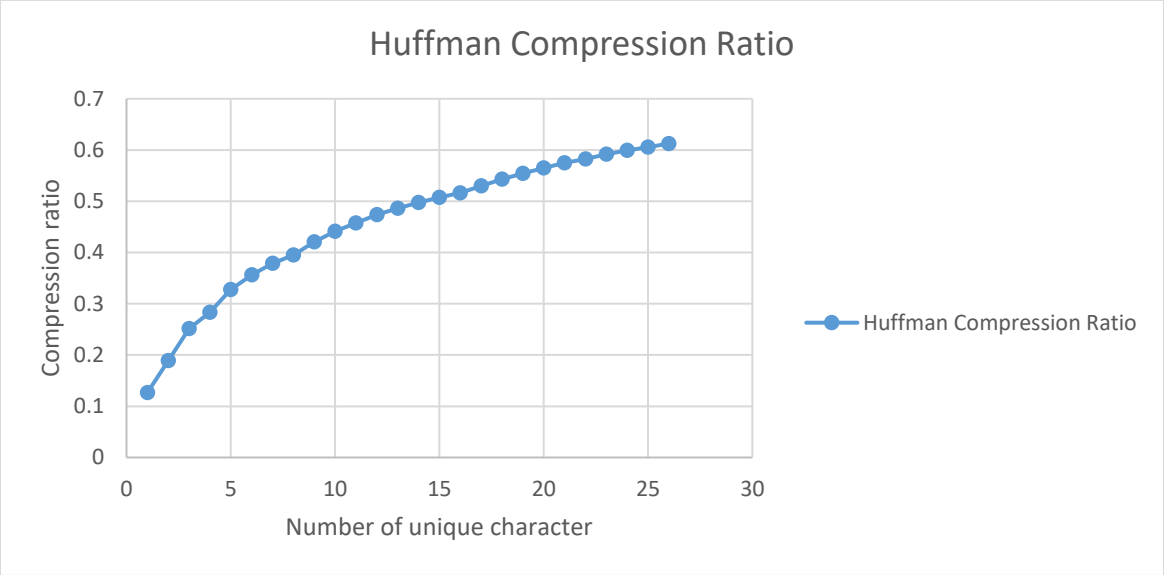
Chenxi Sun

Assignment 12 Document Analysis



1. The way I set up the experiment in graphing the Huffman compression ratio is to first create a helper method generate a string, that has a length of 10000. This is done by using a for loop and appending a char to the string. I also set up a random char generator of 'a'+uniquechar, with uniquechar being of the type int and is a parameter that can be input by the user. The int can be between let say 1 to 26 to change the char from a to z lower case. Once this is done, I create a generatefile helper method which uses printwriter to write the string to a new txt file. This is then used to plot the graph above by doing a for loop and changing the number of unique character from 1 to 26 from the changing of the parameter uniquechar. As seen in the plot above, the compression ratio changed from around .1 for 1 unique character to almost .63 at 26 unique characters. This shows that as the number of the unique character is low the compressed file will be significantly smaller than the decompressed file.
2. The kind of file that will results in lower bits when compressing will be music file, data file with a lot of high frequency characters. The higher the frequency of character as seen in the plot above the better the compression. The kind of file that would be bad would be a text file, books and any file with large amount of unique characters. This is also shown in the plot above as at 26 unique characters the compression ratio is .63.
3. The reason you merge the two smallest tree is that you are using a priority queue. The use of the priority queue is that it pulls out the minimum weighted node so you build the tree by merging the two smallest.
4. The Huffman compression uses lossless compression. Lossless compression is when you compress a file and don't lose any data. This is mainly used in text file or music file or pictures.

For lossy compression, this means you are losing data but only because it is used to save bandwidth. This can be used in large data compression like streaming videos.

5. I spent 5 hours on this assignment, most of it on the analysis.