Jayden Younger

Ming Li

CS 2123 Data Structures

07 November 2024

CS 2123 Programming Project 4

Implementation of my project

I implemented Huffman coding using the following steps:

- Character Frequency Analysis: My implementation of Huffman's coding reads each book text file to count the frequency of each character.
- Building the Huffman Tree: builds the Huffman Tree using a priority queue, assigning longer codes to less frequent characters and shorter codes to more frequent ones to optimize storage space.
- Generating Huffman Codes: After constructing the Huffman Tree, we traversed it to assign unique binary codes to each character.
- Encoding: The process involves substituting each character with its corresponding binary Huffman code, resulting in a lengthy sequence of binary digits. This sequence will then be stored in a file with the .bin extension.
- Calculating Compression Ratio: The compression ratio will be determined by analyzing the sizes of both the compressed and the original text files, allowing for a comparison that reveals how much space has been saved through the compression process. We will attempt to achieve compression ratios between 50% and 75%

- Decoding: We carefully decoded the binary sequence by utilizing the Huffman Tree method, which allowed us to efficiently translate the encoded data back into its original format. After completing the decoding process, we ensured that the output was properly formatted and saved the results as a .txt file for future reference and accessibility.

compression ratios

File Names	Compression Ratios
book1.txt	0.6035795345615564
book2.txt	0.5949025738576526
book3.txt	0.5521984458715145
book4.txt	0.5896459697338537
book5.txt	0.5874337033803169
book6.txt	0.6010194101784445

Huffman coding can achieve compression ratios between 50% and 75%, depending on the content of the text and the frequency distribution of its characters. Our compression ratios range from a low of 0.55 to a high of 0.6, which indicates that these ratios fall well within the expected range and do not exceed these limits.

top 5 most frequent in book1

Symbol	Frequency	Code
Space	24842	110
E	13648	1110
T	10237	1000
O	9326	0110
A	8309	0100