Data Structures (CSC 202-13-2234) - Lab 7

Gregory Leathrum

June 9, 2023

1 Songs

1.1 Gangsta's Paradise

For a rap song, I chose Coolio's "Gangsta's Paradise", we had this frequency encoding:
{'A': 156, 'S': 136, ' ': 493, 'I': 179, 'W': 43, 'L': 78, 'K': 29, 'T': 184, 'H': 100, 'R': 76, 'O': 126, 'U': 46, 'G': 50, 'E': 247, 'V': 34, 'Y': 49, 'F': 21, 'D': 69, 'M': 61, 'N': 172, 'Z': 1, '"': 80, 'C': 22, 'B': 28, ',': 34, 'P': 38, '?': 5, '2': 2, '3': 1, '4': 1}
The song has 2,561 characters, and the compression ratio was approximately 47.5%. Once everything was encoded in binary, we had this:
{'E': '0000', 'M': '000100', 'F': '00010100', '?': '000101010', '3': '000101011000', '4': '000101011001', 'Z': '00010101101', '2': '0001010111', 'K': '0001011', 'H': '00011', 'G': '001000', 'B': '0010010', 'C': '0010011', 'Y': '001010', 'U': '001011', 'T': '0011', 'I': '0100', 'N': '0101', 'W': '011000', 'P': '011001', '"': '01101', 'A': '0111', 'L': '10000', 'R': '10001', 'D': '10010', 'V': '100110', ',': '100111', 'S': '1010', 'O': '1011', ' ': '11'}
From this, we can see that the minimum height of the tree is 2 (the space character only uses 2 bits) and the maximum height is 12 (the number "4" uses 12 bits). This is not a surprise since each word is relatively short, so the space character is the most common letter in the song. There is lots of variety in each verse, but the chorus is exactly the same line repeated over and over, which likely helped the compression. By construction, this Huffman tree is not a balanced tree.

1.2 Handy

For a pop song, I chose Weird Al's "Handy" (a parody of Iggy Azalea's "Fancy"), we had this frequency encoding:
{'S': '00000', 'M': '000010', ',': '0000110', 'F': '0000111', 'N': '00010', 'W': '000110', 'C': '000111', 'L': '00100', 'U': '00101', 'H': '00110', 'K': '0011100',

'?': '0011101000', 'X': '0011101001', '-': '0011101010', '9': '0011101011', '(': '00111011', 'G': '001111', 'E': '0100', 'O': '0101', 'Y': '01100', ')': '01101000', 'V': '01101001', 'J': '011010100', '!': '01101010100', 'Ñ': '01101010101', 'È': '01101010110', 'Z': '01101010111', 'Q': '01101011', '"': '011011', 'T': '0111', 'A': '1000', 'R': '1001', 'B': '101000', 'P': '101001', 'D': '10101', 'I': '1011', ' ': '11'}

The song has 2,257 characters, and the compression ratio was approximately 45.4%. Once everything was encoded in binary, we had this:

{'E': '0100', 'M': '000010', 'F': '0000111', '?': '0011101000', '3': '000101011000', '4': '000101011001', 'Z': '01101010111', '2': '0001010111', 'K': '0011100', 'H': '00110', 'G': '001111', 'B': '101000', 'C': '000111', 'Y': '01100', 'U': '00101', 'T': '0111', 'I': '1011', 'N': '00010', 'W': '000110', 'P': '101001', '"': '011011', 'A': '1000', 'L': '00100', 'R': '1001', 'D': '10101', 'V': '01101001', ',': '0000110', 'S': '00000', 'O': '0101', ' ': '11', 'X': '0011101001', '-': '0011101010', '9': '0011101011', '(': '00111011', ')': '01101000', 'J': '011010100', '!': '01101010100', 'Ñ': '01101010101', 'È': '01101010110', 'Q': '01101011'}

From this, we can see that the minimum height of the tree is 2 (the space character only uses 2 bits) and the maximum height is 12 (the number "3" uses 12 bits). This is not a surprise since each word is relatively short, so the space character is the most common letter in the song. There are no repeated chorus parts and only a few repeated lines, which likely hindered the compression. Furthermore, there are several symbols which only appear once in the song (notably the Spanish letters), so that would increase the height of the tree. By construction, this Huffman tree is not a balanced tree.

1.3 Leukocyte

For another pop song, I chose Tim Blais's "Leukocyte" (a parody of BTS's "Dynamite"), we had this frequency encoding:

{'C': 57, 'A': 99, 'U': 61, 'S': 99, 'E': 219, ' ': 448, 'I': 165, 'G': 90, 'O': 138, 'T': 211, 'H': 107, 'R': 68, 'N': 96, 'W': 26, 'M': 36, 'P': 35, 'Y': 61, 'B': 38, 'L': 84, 'D': 54, 'F': 62, 'X': 2, 'K': 22, '"': 5, "'": 3, 'V': 13, ',': 9, '!': 1, 'J': 1, '-': 5, '?': 1, 'Q': 2}

The song has 2,318 characters, and the compression ratio was approximately 47.6%. Once everything was encoded in binary, we had this:

{'T': '0000', 'H': '00010', 'D': '000110', 'W': '0001110', 'V': '00011110', "'": '0001111100', 'J': '000111110100', '?': '000111110101', '!': '00011111011', '"': '000111111', 'A': '00100', 'S': '00101', 'N': '00110', 'G': '00111', 'I': '0100', 'L': '01010', 'K': '0101100', ',': '01011010', '-': '010110110', 'X': '0101101110', 'Q': '0101101111', 'B': '010111', 'M': '011000', 'P': '011001', 'R': '01101', 'O': '0111', 'F': '10000', 'U': '10001', 'Y': '10010', 'C': '10011', 'E': '101', ' ': '11'}

From this, we can see that the minimum height of the tree is 2 (the space character only uses 2 bits) and the maximum height is 12 (the letter "J" uses 12 bits). Each word is relatively long, but the variety of long scientific and Latin-based

words likely hindered the compression. By construction, this Huffman tree is not a balanced tree.

2 Mantras

2.1 Wreck-It Ralph

For a short mantra, I chose a mantra from the movie "Wreck-It Ralph", which says "I'm bad, and that's good. I will never be good, and that's not bad. There's no one I'd rather be than me." We had this frequency encoding:
{'I': 4, '"': 5, 'M': 2, ' ': 21, 'B': 4, 'A': 8, 'D': 7, ',': 2, 'N': 7, 'T': 8, 'H': 5, 'S': 3, 'G': 2, 'O': 7, '.': 3, 'W': 1, 'L': 2, 'E': 9, 'V': 1, 'R': 4}
This mantra has 105 characters, and the compression ratio was approximately 50.5%. Once everything was encoded in binary, we had this:
{'E': '0000', 'A': '0001', 'T': '0010', 'R': '00110', 'L': '001110', 'W': '0011110', 'V': '0011111', 'I': '01000', 'B': '01001', 'D': '0101', 'N': '0110', 'O': '0111', ',': '100000', 'G': '100001', 'S': '10001', '"': '1001', 'H': '1010', '.': '10110', 'M': '10111', ' ': '11'}
From this, we can see that the minimum height of the tree is 2 (the space character only uses 2 bits) and the maximum height is 7 (the letter "V" uses 7 bits). There is a large variety of letters, but their frequency means that even the most infrequent letters can be encoded in fewer than 8 bits. By construction, this Huffman tree is not a balanced tree.

2.2 Vogon Poetry

For a slightly longer "mantra", I chose a poem from the book "Hitchhiker's Guide To The Galaxy" by Douglas Addams. It is referred to as the "3rd worst poetry in the universe", read out by aliens called the Vogons. We had this frequency encoding:
{'O': 29, 'H': 18, ' ': 85, 'F': 6, 'R': 43, 'E': 50, 'D': 22, 'L': 39, 'G': 24, 'U': 27, 'N': 31, 'T': 37, 'B': 12, 'Y': 9, ',': 16, 'M': 16, 'I': 40, 'C': 9, 'A': 27, 'S': 27, 'P': 9, 'J': 4, 'Q': 1, '.': 2, 'W': 6, 'X': 1, 'V': 2, 'K': 2, '"': 1, '!': 1}
This mantra has 596 characters, and the compression ratio was approximately 46%. Once everything was encoded in binary, we had this:
{'E': '0000', 'J': '00010000', '"': '0001000100', '!': '0001000101', 'Q': '0001000110', 'X': '0001000111', 'F': '0001001', 'B': '000101', 'G': '00011', 'R': '0010', 'D': '00110', 'W': '0011100', 'V': '001110100', 'K': '001110101', '.': '00111011', 'Y': '001111', ' ': '010', 'I': '0110', 'L': '0111', 'T': '1000', 'H': '10010', 'C': '100110', 'P': '100111', ',': '10100', 'M': '10101', 'N': '1011', 'O': '1100', 'U': '1101', 'A': '1110', 'S': '1111'}

From this, we can see that the minimum height of the tree is 3 (the space character only uses 3 bits) and the maximum height is 10 (the letter "X" uses 10 bits). There is a large variety of letters and each word is relatively long, which likely hindered the compression and explains why the space character does not appear more frequently. By construction, this Huffman tree is not a balanced tree.

2.3 Rick And Morty Copypasta

For a long mantra, I chose the famous copypasta about the show "Rick And Morty" which starts with "To be fair, you have to have a very high IQ to understand Rick and Morty...". We had this frequency encoding:
{'T': 99, 'O': 80, ' ': 215, 'B': 12, 'E': 116, 'F': 26, 'A': 76, 'I': 83, 'R': 60, ',': 11, 'Y': 34, 'U': 32, 'H': 63, 'V': 14, 'G': 11, 'Q': 3, 'N': 68, 'D': 33, 'S': 75, 'C': 34, 'K': 11, 'M': 17, '.': 13, 'X': 2, 'L': 40, 'W': 18, 'P': 24, '"': 12, '-': 4, ';': 1, 'J': 3, '&': 2, "'": 2, '5': 1, '(': 1, ')': 1}
This mantra has 1,297 characters, and the compression ratio was approximately 45.7%. Once everything was encoded in binary, we had this:
{' ': '000', 'E': '0010', 'T': '0011', '.': '0100000', 'B': '0100001', 'P': '010001', '"': '0100100', '&': '0100101000', "'": '0100101001', 'Q': '010010101', 'J': '010010110', 'X': '010010111', ',': '0100110', 'G': '0100111', 'I': '0101', 'O': '0110', 'L': '01110', 'K': '0111100', '-': '01111010', '(': '0111101100', ')': '0111101101', ';': '0111101110', '5': '0111101111', 'W': '011111', 'A': '1000', 'S': '1001', 'N': '1010', 'Y': '10110', 'C': '10111', 'D': '11000', 'U': '11001', 'H': '1101', 'R': '1110', 'M': '111100', 'V': '111101', 'F': '11111'}
From this, we can see that the minimum height of the tree is 3 (the space character only uses 3 bits) and the maximum height is 10 (the symbol "&" uses 10 bits). There is a large variety of letters and words, which likely hindered the compression. By construction, this Huffman tree is not a balanced tree.