



**INSTITUTO
FEDERAL**

Santa Catarina

Câmpus
São José

Compressão e Descompressão com Algoritmo de Huffman

Sistemas de Comunicação II

Arthur Cadore Matuella Barcella

01 de Dezembro de 2024

Engenharia de Telecomunicações - IFSC-SJ

Sumário

1. Introdução	3
2. Compressão Huffman	3
2.1. Leitura do arquivo	3
2.2. Análise dos caracteres contidos	4
2.3. Calculo do percentual individual	5
2.4. Calculando a PMF (Probability Mass Function)	6
2.5. Indexação dos caracteres	7
2.6. Aplicando codificação sobre o texto	7
2.7. Imprime o arquivo codificado	8
3. Analisando o arquivo codificado	8
4. Decodificando o arquivo	8
4.1. Leitura do arquivo codificado	8
4.2. Decodificação do arquivo por huffman:	9
4.3. Aplicando a indexação inversa:	9
4.4. Exportando decodificado em um arquivo:	10
5. Conclusão	10

1. Introdução

Neste relatório, será apresentado um estudo sobre o algoritmo de Huffman, incluindo a leitura de um arquivo de texto, a contagem de ocorrências de cada caractere, a análise dos caracteres contidos no arquivo, o cálculo do percentual de cada caractere, a criação de uma Função de Massa de Probabilidade (PMF), a aplicação do algoritmo de Huffman sobre o texto, a codificação do texto e a análise do arquivo codificado.

O algoritmo de Huffman é um método de compressão de dados que utiliza a codificação de caracteres para reduzir o tamanho de um arquivo. O algoritmo foi desenvolvido por David A. Huffman em 1952, e é amplamente utilizado em sistemas de comunicação e armazenamento de dados.

2. Compressão Huffman

2.1. Leitura do arquivo

Inicialmente, deve-se coletar todos os caracteres contidos no arquivo de texto. Para isso, é necessário realizar a leitura do arquivo e contar a quantidade de ocorrências de cada caractere. Isso é executado através do código abaixo:

```
1 # Read the file and count the number of occurrences of each character
2 with open("alice.txt", "r", encoding="utf-8") as file:
3     # Ler o arquivo
4     text = file.read()
5
6     # Create a variable to store the characters and their occurrences
7     characters = {}
8
9     for char in text:
10         # Check if the character is on the variable
11         if char in characters:
12             # Increment the character
13             characters[char] += 1
14         else:
15             # Add the character to the variable and set it to 1
16             characters[char] = 1
17
18     # Order the characters
19     sorted_letters = sorted(characters.items())
20
21 # Get the letters and occurrences
22 letters = [letter for letter, occurrences in sorted_letters]
23 occurrences = [occurrences for letter, occurrences in sorted_letters]
24
25 # Print the letters count:
26 # Print the letters count:
27 print("Letters count:", len(letters))
28
29 BFR = math.ceil(math.log2(len(letters)))
30 print("Bits for representation:", BFR)
```

Desta forma, o vetor `letters` contém todos os caracteres contidos no arquivo de texto, enquanto o vetor `occurrences` contém a quantidade de ocorrências de cada caractere.

Também podemos notar que com base na impressão de “letters count”, o arquivo contém 91 caracteres distintos, necessitando de 91 símbolos diferentes para representar cada caractere, portanto, a codificação mais simples precisará de 7 bits ($2^7 = 128$) para representar todos os caracteres.

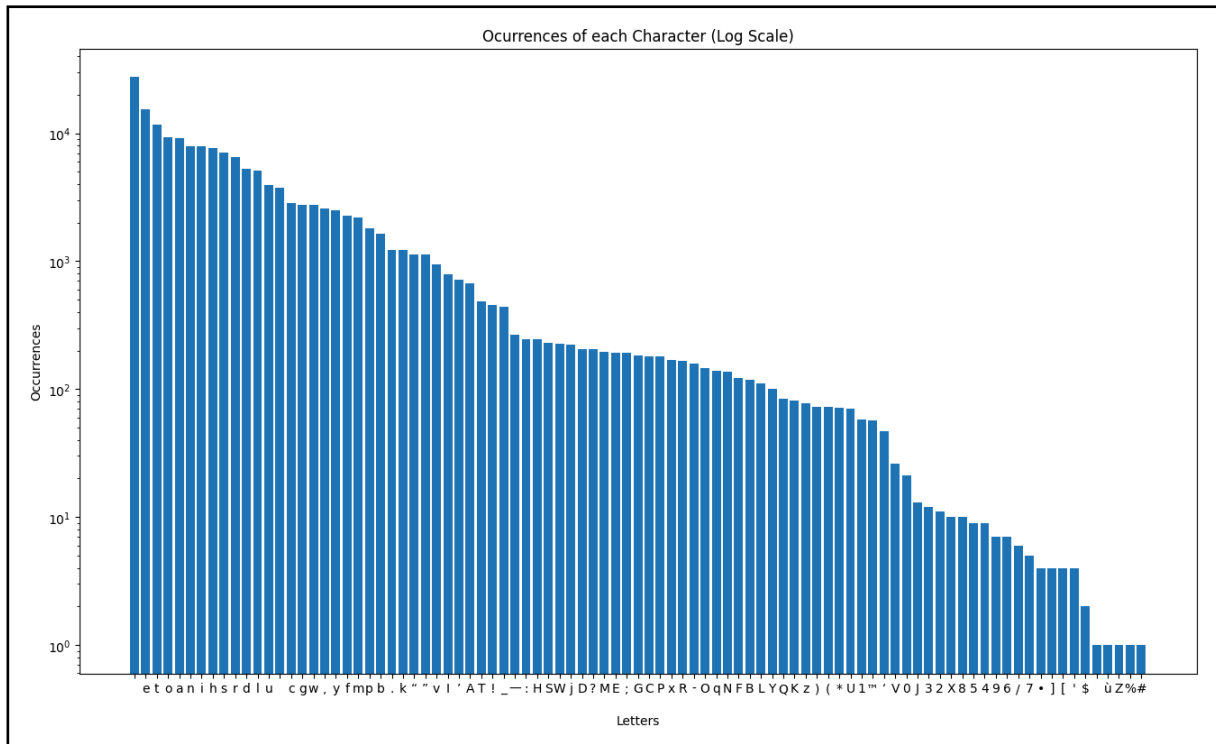
2.2. Análise dos caracteres contidos

Na sequência, é possível analisar a quantidade de ocorrências de cada caractere contido no arquivo de texto. Para isso, foi aplicado um plot de barras, onde o eixo x representa os caracteres e o eixo y representa a quantidade de ocorrências de cada caractere. O código abaixo realiza essa operação:

```
1 sorted_letters = [letter for _, letter in sorted(zip(occurrences, letters),
2 reverse=True)]
3 sorted_occurrences = [occurrence for occurrence, _ in
4 sorted(zip(occurrences, letters), reverse=True)]
5 plt.figure(figsize=(16,9))
6 plt.bar(sorted_letters, sorted_occurrences)
7 plt.yscale("log")
8 plt.xlabel("Letters")
9 plt.ylabel("Occurrences")
10 plt.title("Occurrences of each Character (Log Scale)")
11 plt.show( )
```

Com base neste código, é gerado a seguinte figura:

Figura 1: Elaborada pelo Autor



Note que o primeiro caractere mais frequente é o espaço, seguido pelas letras “e”, “t”, “a” e “o”.

Nota: Após o caracter “u”, aparentemente outro caracter espaço está sendo representado, porém, esse representa o caracter de quebra de linha (“\n”).

2.3. Calculo do percentual individual

Com base na quantidade de ocorrências de cada caractere, é possível calcular o percentual de cada caractere em relação ao total de caracteres contidos no arquivo de texto.

Para isso, calculamos a quantidade total de caracteres lidos, e determinamos a razão entre a quantidade de ocorrências de cada caractere e o total de caracteres. O código abaixo realiza essa operação:

Total de caracteres: 163919

```
1 # sum all the occurrences of the letters and create a graph with the
  # percentage of each letter
2
3 # Get the total number of letters
4 total_letters = sum(occurrences)
5 print("Total letters:", total_letters)
6
7 # Calculate the percentage of each letter
8 def percentage(occurrences, total):
9     return [occurrence / total for occurrence in occurrences]
10
11 percentages = percentage(occurrences, total_letters)
12
13 # sort the letters by occurrences to make the graph more readable
```

Dessa forma é possível visualizar a porcentagem de cada caractere em relação ao total de caracteres contidos no arquivo de texto, conforme apresentado abaixo:

Percentage of each Character (Log Scale)

Character	Percentage (approx.)
e	0.2
t	0.1
o	0.08
a	0.06
n	0.055
i	0.05
h	0.045
s	0.04
r	0.035
d	0.03
l	0.025
u	0.02
c	0.018
g	0.016
w	0.015
y	0.014
f	0.013
m	0.012
p	0.011
b	0.01
k	0.008
"	0.0075
'	0.007
v	0.0065
!	0.006
~	0.0055
_	0.005
H	0.0045
S	0.004
W	0.0038
J	0.0035
D	0.0032
7	0.003
M	0.0028
E	0.0025
G	0.0022
C	0.002
P	0.0018
x	0.0016
R	0.0014
-	0.0012
O	0.001
q	0.0009
N	0.0008
F	0.0007
B	0.00065
L	0.0006
Y	0.00055
Q	0.0005
K	0.00045
Z	0.0004
)	0.00038
(0.00035
*	0.00032
U	0.0003
1	0.00028
"	0.00025
'	0.00022
m	0.0002
"	0.00018
V	0.00015
0	0.00012
J	0.0001
3	8e-05
2	7e-05
X	6.5e-05
8	6e-05
5	5.5e-05
4	5e-05
9	4.5e-05
6	4e-05
/	3.5e-05
7	3e-05
.	2.8e-05
[2.5e-05
'	2.2e-05
\$	2e-05
ü	1.5e-05
Z	1.2e-05
#	1e-05

Com base na lista de caracteres e seus respectivos percentuais, é possível calcular a Função de Massa de Probabilidade (PMF) de cada caractere. Para isso, é necessário criar um dicionário contendo o caractere e seu respectivo percentual. O código abaixo realiza essa operação:

Engenharia de Telecomunicações - IFSC-SJ

```

6 # Calculate the huff code
7 huff = kmm.HuffmanCode(list(pmf.values()))
8
9 # Print the huff code values and the huff ratio
10 print("Huffman code:", huff.codewords)
11 print("Huff Ratio:", huff.rate(list(pmf.values())))
12 print("Compress Ratio:", BFR - huff.rate(list(pmf.values())))

```

Com base nisso, podemos determinar o código de huffman para cada caractere com base em sua participação do percentual total. Além disso, podemos determinar parâmetros resultantes do código de huffman, como o huff ratio e a taxa de compressão:

Huff Ratio: 4.643275 Compress Ratio: 2.356724

Dessa forma, podemos ver que o código aplicado foi capaz de reduzir a quantidade de bits necessária para representar os caracteres do arquivo em 2.35 bits (média).

2.5. Indexação dos caracteres

Em seguida, é necessário criar uma indexação do caractere e sua respectiva letra, de forma que cada caractere seja representado por um índice. O código abaixo realiza essa operação:

```

1 index = {i: letter for i, letter in enumerate(letters)}
2 range(len(letters)):
3     print(letters[i], occurrences[i])
4 print ("Index:", index)

```

2.6. Aplicando codificação sobre o texto

Com base nesta indexação, é possível codificar o texto original, de forma que cada caractere seja representado por um índice. O código abaixo realiza essa operação. De forma que inicialmente o texto é “codificado” pelo index, transformando cada caractere em uma letra, e posteriormente é aplicado o código de huffman sobre o texto codificado.

```

1 # encode the text using the index to create a integer index of the letters
  in the text
2
3 encoded_text = [list(index.keys())[list(index.values()).index(letter)] for
  letter in text]
4
5 print("Encoded text (by index):", encoded_text)
6
7 # create a huffman code for the encoded text
8 huff_encoded = huff.encode(encoded_text)
9
10 print("Huffman encoded text:", huff_encoded)

```

2.7. Imprime o arquivo codificado

Por fim, é possível imprimir o arquivo codificado, de forma que cada caractere seja representado por um índice. O código abaixo realiza essa operação:

```
1 # export the huffman code to a file.
2 with open("huff_encoded.com2", "w") as file:
3     file.write("".join(map(str, huff_encoded)))
```

3. Analisando o arquivo codificado

Por fim, é possível analisar o arquivo codificado, de forma que cada caractere seja representado por um índice. O código abaixo realiza essa operação:

```
1 000000001011011110001011111000111000000100101011101010001000001011101101010
2 10011110110001010111000110101011110000000111010001000011110110000111110101
3 11010110111000101001110000101011101011000000001011110110100101010001111100
4 0011100101100000011011000101101100111000110110110011011011000000001011011
5 100000011110001010100010000111101100000010111110000001101001101111101010
6 000111110000111111001101110110011011111001101101110101101101111100011
7 1111100101101001101111000000010001011110001001110110000000011000100100101
8 00111000011001101101100101001100000001010010000100110111100011000100010010
9 001101000000110000111110100110111101101000000110110101100100101001011100
10 00101001000000010100100110110110001011011000100110110010110110100110000000
11 10100101110000001111100001010000011110001010001001100100010110000101011011
12 011001010000001100011100000010111000110001001011110001011010011001011101
13 0101001000000100001110111000100010101110000000010111100010010010101101001
14 01110110000001100011111101010000111111000100101011011001110001101001101
15 1110100111000111010010000110000111110100110111100010111100011100000010010
16 10111010100010000010111011010101001111011000101011100011010101000000011111
17 [...]
```

4. Decodificando o arquivo

4.1. Leitura do arquivo codificado

Para decodificar o arquivo, é necessário realizar a leitura do arquivo codificado e aplicar o algoritmo de Huffman para decodificar o texto. O código abaixo realiza essa operação:

```
1 # read the huffman code from the file
2
3 with open("huff_encoded.com2", "r") as file:
4     huff_encoded = file.read()
5
6 # decode the huffman code
7 huffDecode = kmm.HuffmanCode(list(pmf.values()))
8 print("Huffman code:", huffDecode.codewords)
```

Com base no código acima, deve-se obter as palavras código utilizadas originalmente para codificar o texto.

4.2. Decodificação do arquivo por huffman:

Tendo as palavras código e o texto codificado, é possível decodificar o texto original. Esse processo é realizado através do código abaixo:

```
1 # Decode the encoded text
2
3 # Convert the string back to a list of integers
4 huff_encoded_list = list(map(int, huff_encoded))
5
6 decoded_text = huffDecode.decode(huff_encoded_list)
7
8 # print more values of the decoded text
9 print("Decoded text:", decoded_text[:100])
```

O texto decodificado é apresentado abaixo, note que ainda não é possível ler pois o texto está estruturado com base em um index de cada letra correspondente.

```
1 90 46 63 60 1 42 73 70 65 60 58 75 1 33 76 75 60 69 57 60 73 62 1 60
2 28 70 70 66 1 70 61 1 27 67 64 58 60 6 74 1 27 59 77 60 69 75 76 73
3 60 74 1 64 69 1 49 70 69 59 60 73 67 56 69 59 0 1 1 1 1 0 46 63
4 64 74 1 60 57 70 70 66 1 64 74 1 61 70 73 1 75 63 60 1 76 74 60 1
5 70 61 1 56
```

4.3. Aplicando a indexação inversa:

Dessa forma, para resolver o problema de leitura do texto decodificado, é necessário aplicar a indexação inversa, de forma que cada índice seja representado por um caractere. Abaixo está o index utilizado no processo de codificação (gerado automaticamente durante a codificação).

```
1 0: '\n', 1: ' ', 2: '!', 3: '#', 4: '$', 5: '%', 6: '"', 7: '(', 8: ')'
2 9: '*', 10: ',', 11: '-', 12: '.', 13: '/', 14: '0', 15: '1', 16: '2',
3 17: '3', 18: '4', 19: '5', 20: '6', 21: '7', 22: '8', 23: '9', 24: ':',
4 25: ';', 26: '?', 27: 'A', 28: 'B', 29: 'C', 30: 'D', 31: 'E', 32: 'F',
5 33: 'G', 34: 'H', 35: 'I', 36: 'J', 37: 'K', 38: 'L', 39: 'M', 40: 'N',
6 41: 'O', 42: 'P', 43: 'Q', 44: 'R', 45: 'S', 46: 'T', 47: 'U', 48: 'V',
7 49: 'W', 50: 'X', 51: 'Y', 52: 'Z', 53: '[', 54: ']', 55: '_', 56: 'a',
8 57: 'b', 58: 'c', 59: 'd', 60: 'e', 61: 'f', 62: 'g', 63: 'h', 64: 'i',
9 65: 'j', 66: 'k', 67: 'l', 68: 'm', 69: 'n', 70: 'o', 71: 'p', 72: 'q',
10 73: 'r', 74: 's', 75: 't', 76: 'u', 77: 'v', 78: 'w', 79: 'x', 80: 'y',
11 81: 'z', 82: 'ù', 83: '-', 84: "'", 85: '"', 86: '"', 87: '"', 88: '•',
12 89: '™', 90: '\uffff'}
```

Esse index é aplicado no código abaixo para realizar a indexação inversa do texto, retornando-o em palavras.

```
1 # use the index to convert the integers back to letters, the index need to
2   check all the values of the index to find the letter
3 decoded_text = [index[i] for i in decoded_text]
4
```

```
5 print("Decoded text:", decoded_text[:100])
```

4.4. Exportando decodificado em um arquivo:

Por fim, é possível exportar o texto decodificado para um arquivo de texto. O código abaixo realiza essa operação:

```
1 # print the text into a file "huff_decoded.txt"
2
3 with open("huff_decoded.txt", "w") as file:
4     file.write("".join(decoded_text))
```

5. Conclusão

Neste relatório, foi apresentado um estudo sobre o algoritmo de Huffman, incluindo a leitura de um arquivo de texto, a contagem de ocorrências de cada caractere, a análise dos caracteres contidos no arquivo, o cálculo do percentual de cada caractere, a criação de uma Função de Massa de Probabilidade (PMF), a aplicação do algoritmo de Huffman sobre o texto, a codificação do texto e a análise do arquivo codificado.

O algoritmo de Huffman é um método de compressão de dados que utiliza a codificação de caracteres para reduzir o tamanho de um arquivo. O algoritmo foi desenvolvido por David A. Huffman em 1952, e é amplamente utilizado em sistemas de comunicação e armazenamento de dados.