Huffman Coding Compression Algorithm

Elif Özbay

18.11.2020

1 Occurences

The percentages of the letters in the text is calculated as in Figure 1.

Letter	Percentage
a	6.30%
b	1.36%
c	2.27%
d	3.00%
e	10.44%
f	2.00%
g	1.35%
h	4.73%
i	6.34%
j	0.06%
k	0.42%
l	2.64%
m	2.03%
n	6.37%
o	6.47%
p	1.41%
q	0.10%
r	4.48%
s	5.65%
t	8.03%
u	2.14%
v	0.79%
w	1.62%
x	0.23%
y	1.36%
z	0.02%
' '	18.35%

Table 1: Letter Occurences

2 Compression Size Metrics

	Original File	Filtered File	Compressed File	Extracted File
Size (b)	1,628,965	1,550,801	988,512	1,550,801
Size (mb)	1.62	1.55	0.988	1.55

Table 2: Size Metrics

3 Lookup Table

Letter	Code
a	1000
b	011100
c	00001
d	01111
e	010
f	110010
g	001110
h	0010
i	1001
j	11000110001
k	11000111
l	00110
m	110011
n	1010
o	1011
p	011101
q	1100011001
r	0001
s	0110
t	1101
u	00000
v	1100010
w	110000
x	110001101
y	001111
z	11000110000
' '	111

Table 3: Codings

4 Entropy

$$H = -\sum_{i=1}^{27} p_i log_2 p_i$$

$$\begin{aligned} = -0.063 * log_2 0.063 - 0.0136 * log_2 0.0136 - 0.0227 * log_2 0.0227 - 0.03 * log_2 0.03 \\ - 0.1044 * log_2 0.1044 - 0.02 * log_2 0.02 - 0.0135 * log_2 0.0135 - 0.0473 * log_2 0.0473 \\ - 0.0634 * log_2 0.0634 - 0.0006 * log_2 0.0006 - 0.0042 * log_2 0.0042 - 0.0264 * log_2 0.0264 \\ - 0.0203 * log_2 0.0203 - 0.0637 * log_2 0.0637 - 0.0647 * log_2 0.0647 - 0.0141 * log_2 0.0141 \\ - 0.001 * log_2 0.001 - 0.0448 * log_2 0.0448 - 0.0565 * log_2 0.0565 - 0.0803 * log_2 0.0803 \\ - 0.0214 * log_2 0.0214 - 0.0079 * log_2 0.0079 - 0.0162 * log_2 0.0162 - 0.0023 * log_2 0.0023 \\ - 0.0136 * log_2 0.0136 - 0.0002 * log_2 0.0002 - 0.1835 * log_2 0.1835 \end{aligned}$$

$$H = 4.0594 \text{ bits}$$

5 Efficiency

Letter	Original	Compressed
a	8 bits	4 bits
b	8 bits	6 bits
c	8 bits	5 bits
d	8 bits	5 bits
e	8 bits	3 bits
f	8 bits	6 bits
g	8 bits	6 bits
h	8 bits	4 bits
i	8 bits	4 bits
j	8 bits	11 bits
k	8 bits	8 bits
l	8 bits	5 bits
m	8 bits	6 bits
n	8 bits	4 bits
o	8 bits	4 bits
p	8 bits	6 bits
q	8 bits	10 bits
r	8 bits	4 bits
s	8 bits	4 bits
t	8 bits	4 bits
u	8 bits	5 bits
v	8 bits	7 bits
w	8 bits	6 bits
x	8 bits	9 bits
y	8 bits	6 bits
z	8 bits	11 bits
' '	8 bits	3 bits

Size Formula

$$size = \sum_{i=1}^{27} l_i * k_i$$

where,

l_i = code length of character i,

k_i = # occurences of character i

t = total # characters = 1550801

Original Text

$$\begin{aligned} \text{size} &= 8 * t \\ &= 8 * 1550801 \\ &= 12406408 \text{ bits} \\ &= 1550801 \text{ bytes} \end{aligned}$$

Compressed Text

$$\begin{aligned} \text{size} &= 4 * 0.063 + 6 * 0.0136 + ... \\ &= 6353941 \text{ bits} \\ &= 794242 \text{ bytes} \end{aligned}$$

Efficiency

$$\begin{aligned} &\frac{794242}{1550801} \\ &= 0.512 \end{aligned}$$