Trabajo Laboratorio Introducción Sistemas Multimedia II (Práctica Data Compression - Implementación Código Huffman)

En esta página web: https://es.wikipedia.org/wiki/Frecuencia_de_aparici%C3%B3n_de_letras, encontrarás la frecuencia de aparición de letras en español.

A partir de esos valores realiza el código Huffman correspondiente (código Huffman español).

Calcula además la entropía (entropía del español) y compárala con el promedio en longitud de la codificación realizada.

En el documento explicas la naturaleza de la información de la que partes y la forma en la que realizas el código de Huffman. Incluye con screenshots, los contenidos de las tablas creadas (bien dibujadas a mano o bien implementadas en una hoja de cálculo).

Incluye un breve párrafo en el que explicas tus conclusiones.

Vídeo: graba un video de 2 o 3 minutos de duración en el que muestras el desarrollo de tu trabajo.

El Código de Huffman

El código de Huffman es un algoritmo utilizado para la compresión de datos. Funciona asignando códigos binarios de longitud variable a símbolos, de manera que los símbolos más frecuentes tengan códigos más cortos, mientras que los menos frecuentes tienen códigos más largos. Esto permite una representación más eficiente de los datos, reduciendo así el tamaño total del archivo.

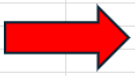
Algoritmo de Huffman

Para empezar con el algoritmo de Huffman se toman los datos de frecuencia de aparición de letras en español y se realiza una tabla con ellas.

LETRA	PORCETAJE (%)
A	12,53
B	1,42
C	4,68
D	5,86
E	13,68
F	0,69
G	1,01
H	0,7
I	6,25
J	0,44
K	0,02
L	4,97
M	3,15
N	6,71
Ñ	0,31
O	8,68
P	2,51
Q	0,88
R	6,87
S	7,98
T	4,63
U	3,93
V	0,9
W	0,01
X	0,22
Y	0,9
Z	0,52
TOTAL	100,45

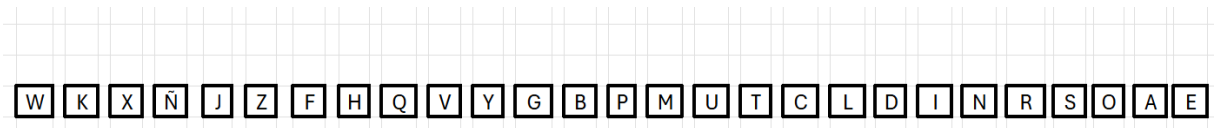
Posteriormente se debe reescribir la tabla ordenando los elementos de menor a mayor por su frecuencia de aparición.

LETRA	PORCETAJE (%)
A	12,53
B	1,42
C	4,68
D	5,86
E	13,68
F	0,69
G	1,01
H	0,7
I	6,25
J	0,44
K	0,02
L	4,97
M	3,15
N	6,71
Ñ	0,31
O	8,68
P	2,51
Q	0,88
R	6,87
S	7,98
T	4,63
U	3,93
V	0,9
W	0,01
X	0,22
Y	0,9
Z	0,52
TOTAL	100,45

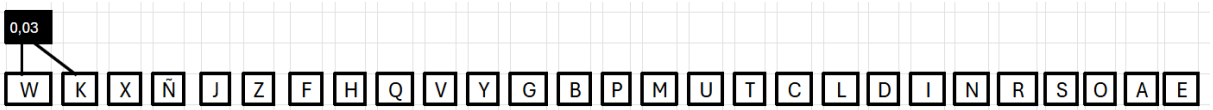


LETRA	PORCETAJE (%)
W	0,01
K	0,02
X	0,22
Ñ	0,31
J	0,44
Z	0,52
F	0,69
H	0,7
Q	0,88
V	0,9
Y	0,9
G	1,01
B	1,42
P	2,51
M	3,15
U	3,93
T	4,63
C	4,68
L	4,97
D	5,86
I	6,25
N	6,71
R	6,87
S	7,98
O	8,68
A	12,53
E	13,68
TOTAL	100,45

Seguidamente, para realizar el árbol, es necesario poner todos los nodos ordenados por su frecuencia, a esto se le conoce como el Bosque Inicial.



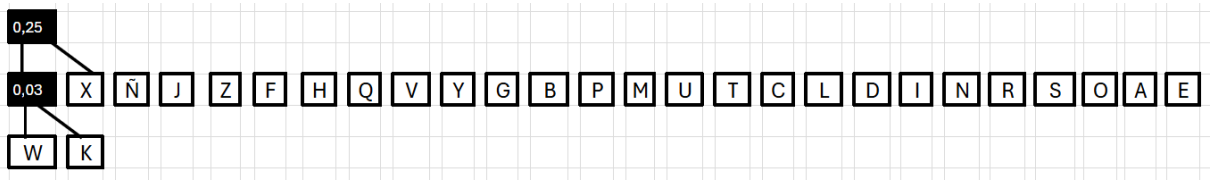
El siguiente paso consiste en crear un nuevo nodo que enlace los dos nodos de menor frecuencia y poner la suma de sus frecuencias.



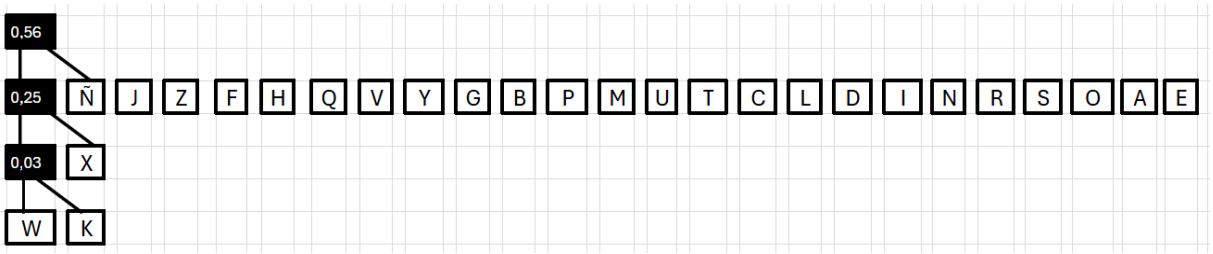
Ahora daremos paso a una serie de iteraciones para construir el árbol.

Lo primero que se debe hacer es reordenar los nodos por su aparición de frecuencia, en este caso la X, que es el elemento posterior, tiene una frecuencia mayor al nodo que hemos creado con anterioridad por lo que el orden sigue manteniéndose.

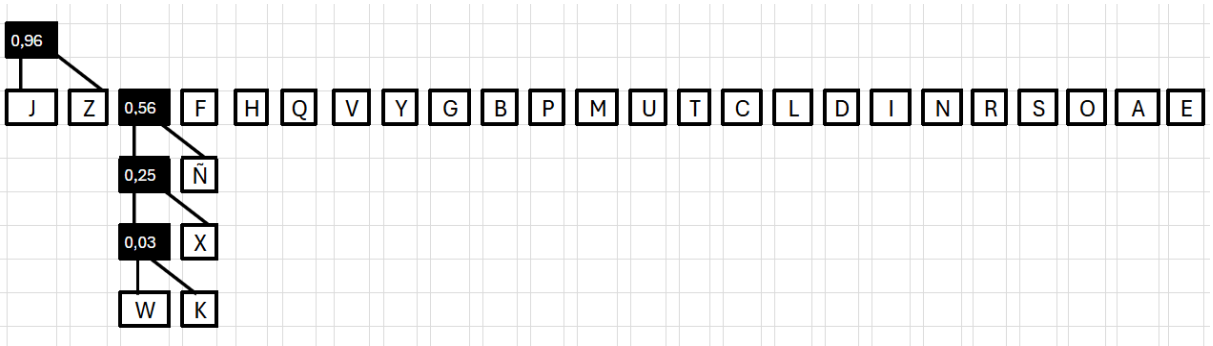
Ahora se creará un nuevo nodo de la misma forma que se ha realizado anteriormente. Se unirán los dos nodos de menor frecuencia que en este caso serán el nodo X y el nodo que se había creado.



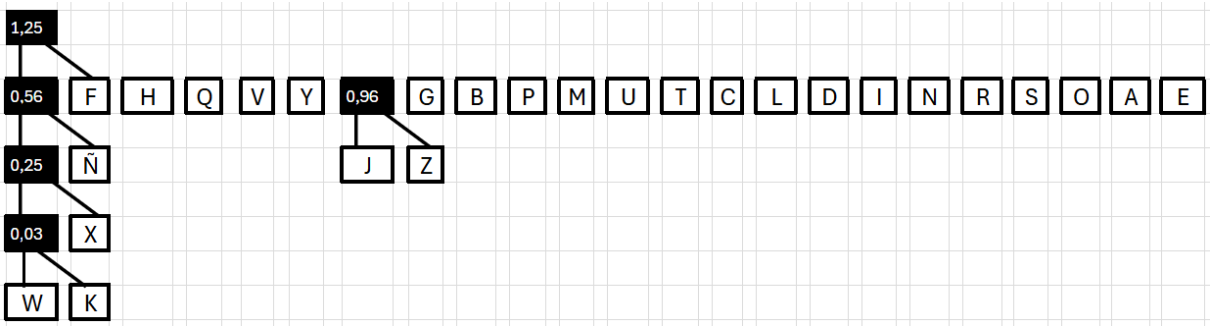
En la siguiente iteración el orden se sigue manteniendo creándose un nodo con la Ñ y el nodo del paso anterior.



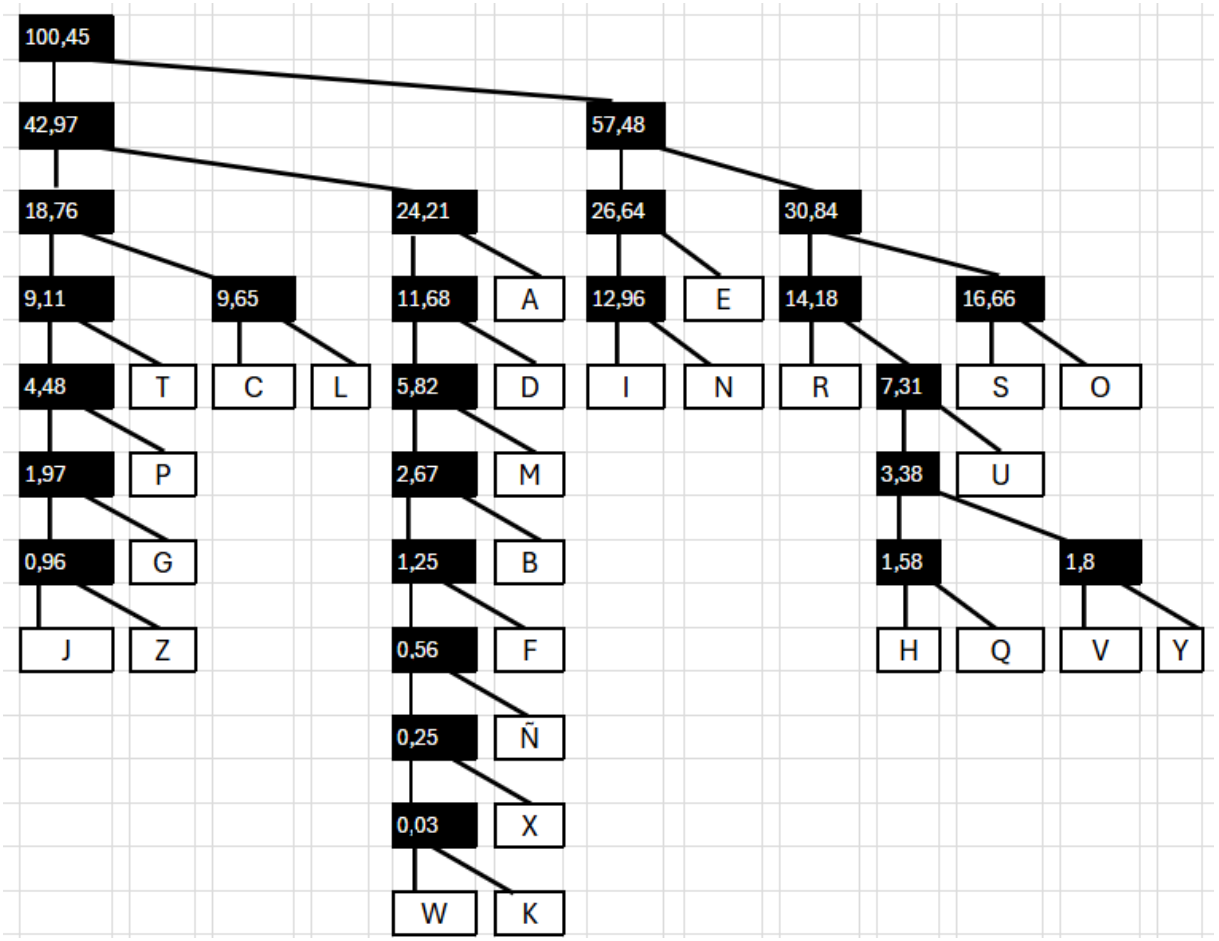
En la tercera iteración se observa que la J y la Z tienen una frecuencia menor a la suma del último nodo por lo que pasan a ponerse delante creando un nuevo nodo entre ambas.



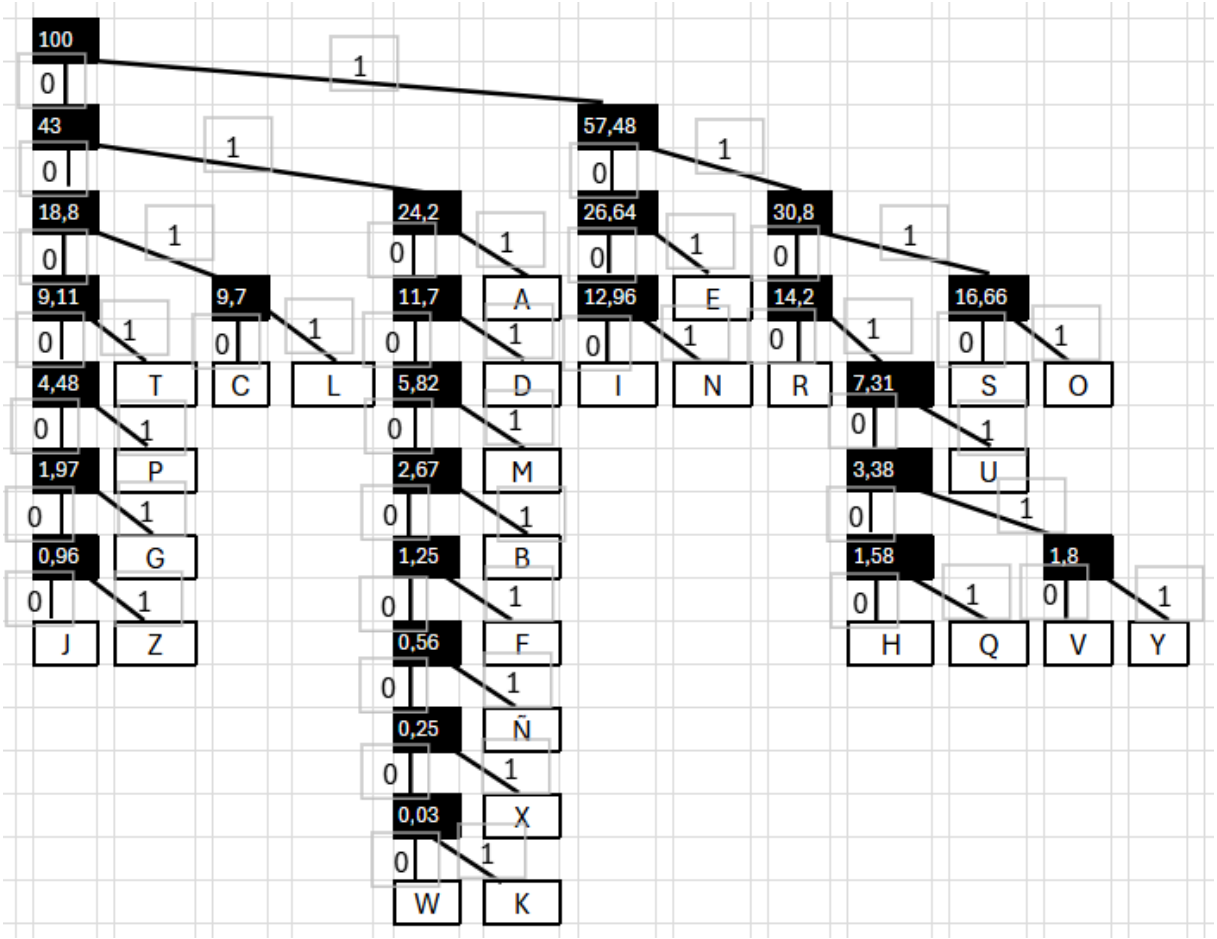
Para la cuarta iteración se vuelve a ordenar quedando el nodo recién creado después de varios elementos ya que la suma resultante es mayor a la de la frecuencia de aparición de dichos elementos.



Se realizan las iteraciones necesarias hasta que todos los elementos terminan confluyendo en un solo nodo.

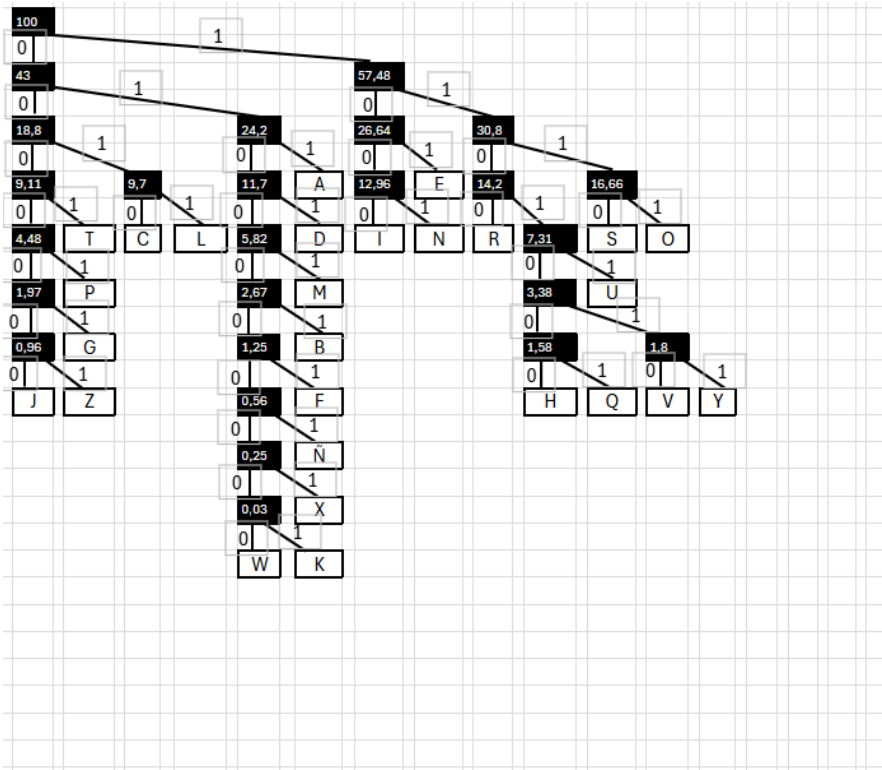


Una vez que se ha construido el árbol asignaremos a las aristas derechas de los nodos con un 1 y a las aristas izquierdas con un 0.



Finalmente, para obtener el código de Huffman de las letras, recorremos las ramas que se atraviesan desde el vértice superior del árbol hasta las letras. Por ejemplo, la letra E, que es la de mayor frecuencia, atraviesa un total de 3 ramas teniendo el código de 101. Tomando el ejemplo de la W, letra con menor frecuencia, se deben atravesar 10 ramas para llegar a ella teniendo como código 0100000000.

Se completa la tabla obteniendo un código específico para cada letra.



LETRA	CÓDIGO
A	011
B	010001
C	0010
D	0101
E	101
F	0100001
G	000001
H	1101000
I	1000
J	0000000
K	0100000001
L	0011
M	01001
N	1001
Ñ	01000001
O	1111
P	00001
Q	1101001
R	1100
S	1110
T	0001
U	11011
V	1101010
W	0100000000
X	010000001
Y	1101011
Z	0000001

Longitud promedio

La longitud promedio es una medida del número promedio de bits necesarios para codificar cada símbolo de un conjunto de datos. Para calcularla se tiene que realizar la siguiente fórmula que donde p es la probabilidad de aparición de la letra y la l es la longitud del código en bits.

$$L_{promedio} = \sum_{i=1}^{n} p_i \cdot l_i$$

Se realiza el producto para cada letra y finalmente se suma todos los elementos consiguiendo la longitud promedio que es de 4,0765 bits.

LETRA	PROBABILIDAD	Pr DECIMAL	CÓDIGO	LONGITUD CÓDIGO	p*l
A	12,53	0,1253	011	3	0,3759
B	1,42	0,0142	010001	6	0,0852
C	4,68	0,0468	0010	4	0,1872
D	5,86	0,0586	0101	4	0,2344
E	13,68	0,1368	101	3	0,4104
F	0,69	0,0069	0100001	7	0,0483
G	1,01	0,0101	000001	6	0,0606
H	0,7	0,007	1101000	7	0,049
I	6,25	0,0625	1000	4	0,25
J	0,44	0,0044	0000000	7	0,0308
K	0,02	0,0002	0100000001	10	0,002
L	4,97	0,0497	0011	4	0,1988
M	3,15	0,0315	01001	5	0,1575
N	6,71	0,0671	1001	4	0,2684
Ñ	0,31	0,0031	01000001	8	0,0248
O	8,68	0,0868	1111	4	0,3472
P	2,51	0,0251	00001	5	0,1255
Q	0,88	0,0088	1101001	7	0,0616
R	6,87	0,0687	1100	4	0,2748
S	7,98	0,0798	1110	4	0,3192
T	4,63	0,0463	0001	4	0,1852
U	3,93	0,0393	11011	5	0,1965
V	0,9	0,009	1101010	7	0,063
W	0,01	0,0001	0100000000	10	0,001
X	0,22	0,0022	010000001	9	0,0198
Y	0,9	0,009	1101011	7	0,063
Z	0,52	0,0052	0000001	7	0,0364
					4,0765

Entropía

La entropía de una fuente de información mide la cantidad promedio de información que se obtiene al observar un símbolo generado por la fuente. Se calcula usando la fórmula de la entropía de Shannon donde p es la probabilidad de aparición del símbolo.

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Se calcula la entropía de cada letra y finalmente se suma todos los elementos consiguiendo la entropía total que es de 4,0423 bits. Este valor indica cuántos bits de información en promedio se requieren para codificar cada letra en el idioma español si se usara una codificación óptima.

LETRA	PROBABILIDAD	Pr DECIMAL	log2(p)	p*log2(p)
A	12,53	0,1253	-2,9965417	-0,3754667
B	1,42	0,0142	-6,1379653	-0,0871591
C	4,68	0,0468	-4,4173477	-0,2067319
D	5,86	0,0586	-4,0929555	-0,2398472
E	13,68	0,1368	-2,8698599	-0,3925968
F	0,69	0,0069	-7,1791879	-0,0495364
G	1,01	0,0101	-6,6295009	-0,066958
H	0,7	0,007	-7,1584294	-0,050109
I	6,25	0,0625	-4	-0,25
J	0,44	0,0044	-7,8282808	-0,0344444
K	0,02	0,0002	-12,287712	-0,0024575
L	4,97	0,0497	-4,3306103	-0,2152313
M	3,15	0,0315	-4,9885044	-0,1571379
N	6,71	0,0671	-3,8975434	-0,2615252
Ñ	0,31	0,0031	-8,3335161	-0,0258339
O	8,68	0,0868	-3,5261611	-0,3060708
P	2,51	0,0251	-5,3161688	-0,1334358
Q	0,88	0,0088	-6,8282808	-0,0600889
R	6,87	0,0687	-3,8635461	-0,2654256
S	7,98	0,0798	-3,6474674	-0,2910679
T	4,63	0,0463	-4,432844	-0,2052407
U	3,93	0,0393	-4,6693269	-0,1835045
V	0,9	0,009	-6,7958593	-0,0611627
W	0,01	0,0001	-13,287712	-0,0013288
X	0,22	0,0022	-8,8282808	-0,0194222
Y	0,9	0,009	-6,7958593	-0,0611627
Z	0,52	0,0052	-7,5872727	-0,0394538
				4,04239981

Conclusión

La longitud promedio de un código es una medida que indica cuántos bits en promedio se necesitan para codificar cada símbolo de un conjunto de datos. En este caso, el cálculo de la longitud promedio, utilizando la fórmula que combina la probabilidad de aparición de cada símbolo y la longitud del código en bits, da como resultado un valor de 4,0765 bits.

Por otro lado, la entropía es una medida de la cantidad promedio de información proporcionada por un símbolo en una fuente de datos. La entropía total calculada, de acuerdo con la fórmula de Shannon, es de 4,0423 bits, lo que refleja la cantidad de bits necesarios para codificar cada símbolo de manera óptima.

Al comparar ambos valores, la longitud promedio es ligeramente mayor que la entropía, lo que indica que la codificación utilizada es cercana a la óptima, pero no completamente eficiente, ya que se requieren más bits de los estrictamente necesarios según la entropía.