# Código Huffman y primer teorema de Shannon

## 1. Introducción

El árbol de Huffman es un algoritmo de compresión de datos sin pérdida que utiliza una estructura de árbol binario para asignar códigos de longitud variable a cada símbolo en un conjunto de datos.

En términos matemáticos, el árbol de Huffman se basa en la probabilidad de aparición de cada símbolo en el conjunto de datos y utiliza esta información para construir un árbol binario óptimo.

El algoritmo comienza creando nodos para cada símbolo y asignando una probabilidad a cada uno. Luego, combina los dos símbolos con las probabilidades más bajas en un nuevo nodo, cuya probabilidad es la suma de las probabilidades de los nodos originales.

Este proceso se repite hasta que todos los nodos se combinan en un único nodo raíz que representa todo el conjunto de datos. Los códigos de longitud variable se asignan a cada símbolo en función de su posición en el árbol: los símbolos que aparecen con más frecuencia tienen códigos más cortos, mientras que los símbolos menos comunes tienen códigos más largos.

En resumen, el árbol de Huffman es un algoritmo matemático que utiliza la probabilidad de los símbolos en un conjunto de datos para construir un árbol binario óptimo que asigna códigos de longitud variable a cada símbolo.

## 2. Arquitectura

En nuestro caso particular, la respresentación del árbol será una tupla de tuplas. Definimos el árbol por inducción en su 'complejidad', siendo $a_n$ un árbol de complejidad $n$:

- Caso base $n = 0$:
$$a_0 = (x, \omega),$$
donde $a_0$ representa una hoja del árbol de huffman con la letra $x$ y peso $\omega \in [0, 1]$.
- Caso inductivo $n > 0$:
$$a_n = ((a^*_{k_1}, a^*_{k_2}), \omega^1 + \omega^2),$$
donde $k_1, k_2 < n$ y además $a^*_{k_1}$ y $a^*_{k_2}$ son los hijos izquierdo y derecho respectivamente de $a_n$, siendo $a^*_{k_i} = a_{k_i}[0]$ (la primera componente de $a_{k_i}$ obviando así el peso), para $i = 1, 2$. Y $\omega^i$ es el peso del árbol $a_{k_i}$ para $i = 1, 2$.

## 3. Material usado

Para esta práctica se han usado dos archivos de texto externos *español.txt* y *ingles.txt*. La práctica consiste en hacer una codificación huffman de dichos archivos de texto y de ahí obtener ciertos resultados.

# 4. Resultados

## 4.1. Pregunta 1

**Español**

- Longitud media:

$$L(C) = \frac{1}{W}\sum_{i=1}^{n} w_i|x_i| = 4{,}37$$

- Comprobación del primer teorema de Shannon:

$$H(C) = -\sum_{i=1}^{n} p_i \log_2(p_i) = 4{,}34$$

por tanto

$$H(C) \leq L(C) < H(C) + 1$$

- Codigos del alfabeto:

$a \rightarrow 010$
á $\rightarrow 0111110$
$b \rightarrow 0110111$
$c \rightarrow 11100$
$d \rightarrow 11110$
$e \rightarrow 000$
é $\rightarrow 1110101000$
$f \rightarrow 0111101$
$g \rightarrow 1011000$
$h \rightarrow 0111111$
$i \rightarrow 0011$
í $\rightarrow 01101101$
$j \rightarrow 0110010$
$l \rightarrow 0010$
$m \rightarrow 111011$
$n \rightarrow 1001$
$o \rightarrow 1000$
ó $\rightarrow 0110100$
$p \rightarrow 101101$
$q \rightarrow 11101011$
$r \rightarrow 11111$
$s \rightarrow 1010$
$t \rightarrow 01110$
$u \rightarrow 10111$
$v \rightarrow 01100011$
$x \rightarrow 1110101001$
$y \rightarrow 0110000$
$z \rightarrow 01101011$
. $\rightarrow 0111100$
, $\rightarrow 0110011$
( $\rightarrow 1110101011$
) $\rightarrow 1110101010$

**Inglés**

- Longitud media:

$$L(C) = \frac{1}{W}\sum_{i=1}^{n} w_i|x_i| = 4{,}44$$

- Comprobación del primer teorema de Shannon:

$$H(C) = -\sum_{i=1}^{n} p_i \log_2(p_i) = 4{,}41$$

por tanto

$$H(C) \leq L(C) < H(C) + 1$$

- Codigos del alfabeto:

$a \rightarrow 1011$
$b \rightarrow 10010010$
$c \rightarrow 01111$
$d \rightarrow 01001$
$e \rightarrow 000$
$f \rightarrow 111011$
$g \rightarrow 100110$
$h \rightarrow 11111$
$i \rightarrow 0101$
$j \rightarrow 001000101$
$k \rightarrow 10011110$
$l \rightarrow 11100$
$m \rightarrow 01110$
$n \rightarrow 1000$
$o \rightarrow 0110$
$p \rightarrow 00101$
$q \rightarrow 001000111$
$r \rightarrow 11110$
$s \rightarrow 0011$
$t \rightarrow 1010$
$u \rightarrow 01000$
$v \rightarrow 10011111$
$w \rightarrow 001001$
$x \rightarrow 111010100$
$y \rightarrow 1001110$
$z \rightarrow 111010011$
. $\rightarrow 1001011$
, $\rightarrow 1001000$
( $\rightarrow 001000011$
) $\rightarrow 001000010$

### 4.2. Pregunta 2

| **Español** | **Inglés** |
|---|---|
| ▪ Codificacion de *dimension*: | ▪ Codificacion de *dimension*: |
| 11110001111101100010011010001110001001 | 01001011001110000100000110110010110000 |
| ▪ Longitud: 38 | ▪ Longitud: 37 |
| ▪ Longitud ascii: 72 | ▪ Longitud ascii: 72 |

### 4.3. Pregunta 3

Sea

$$cod = 011000110101011100101111100010111111010110001101110$$

entonces, la decodificación (del texto en inglés) de dicho código es

$$\text{decodificar}(cod) = isomorphism$$

## 5. Conclusión

Primero de todo destacar que el primer teorema de Shannon se cumple, como era de esperar. El hecho de poner únicamente 2 decimales es debido a que el error de la entropía para el español y el inglés respectivamente son:

$$E_{es}(C) = \sqrt{\Delta^2 \cdot \sum_{i=1}^{n} \left( \log_2(p_i) + \frac{1}{\ln 2} \right)^2} = 0{,}043$$

$$E_{en}(C) = \sqrt{\Delta^2 \cdot \sum_{i=1}^{n} \left( \log_2(p_i) + \frac{1}{\ln 2} \right)^2} = 0{,}047$$

Además, podemos observar en la primera pregunta las longitudes medias ponderadas de los caracteres en los textos español e inglés son repectivamente 4.37 y 4.44, bastante menores que 8 (la longitud usual con codificación ascii). Hay que tener en cuenta que estas medias son ponderadas, es decir, que unas letras pesas más que otras en función de la frecuencia que aparecen en el texto. Esto es precisamente un punto fuerte en la codificación con un árbol de huffman, pues se asegura de que los caracteres que vayan a ser más utilizados tengan una codificación más corta que los que aparezcan menos.

También, mediante un pequeño ejemplo, se puede apreciar en la pregunta 2 como la proporción entre la longitud de nuestro código frente al código ascii de la palabra *dimension* es similar a la total del texto:

- Español:

$$P_{dimension} = \frac{38}{72} = 0{,}527 \simeq 0{,}546 = \frac{4{,}3713}{8} = P_{total}$$

- Inglés:

$$P_{dimension} = \frac{37}{72} = 0{,}514 \simeq 0{,}554 = \frac{4{,}4370}{8} = P_{total}$$