Compresión de datos

Ing. Max Alejandro Antonio Cerna Flores

Agenda

- Definición y Propósito de la compresión de datos
- Algoritmo RLE
- Compresión por probabilidad
- Definición de Entropía y Redundancia
- Características de la entropía de datos.
- La entropía de Shannon

Definición y Propósito

“La compresión de datos es una reducción en el número de bits necesarios para representar los datos”.

La compresión de datos puede ahorrar capacidad de almacenamiento, acelerar la transferencia de archivos y reducir los costos de hardware de almacenamiento y ancho de banda de la red.

Cualquier compresión en particular es con pérdida o sin pérdida.

Es posible implementar la teoría de árboles a la compresión de datos.

Definición y Propósito

Compresión con pérdida: los datos antes y después de comprimirlos son exactos y no se pierde ningún valor.

Compresión sin pérdida: puede eliminar datos para disminuir aún más el tamaño, con lo que reduce la calidad.

Algoritmo RLE (Run Length Encoding)

Simple de compresión de datos sin pérdidas que se ejecuta en secuencias con el mismo valor que ocurre muchas veces consecutivas.

Ejemplo: WWWWWWWWWWWWWWBWWWWWWWWWWWWWWBBBWWWWWWWWWWWWWWWWWWWWWWWWWWWWBWWWWWWWWWWWWWW

Salida: 12W1B12W3B24W1B14W

NOTA: Tener en cuenta que el tamaño de salida se puede duplicar de tamaño en el peor de los casos. (ej.)

ABCD —> A1B1C1D1

Compresión por probabilidad

Alfabeto de tamaño ***n***

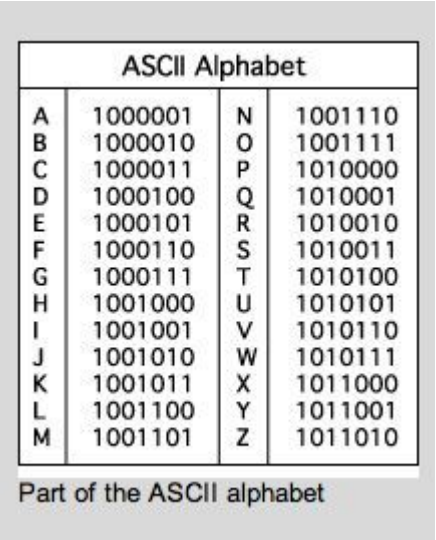
Número de bits necesarios para representar cada símbolo: ***log2(n)***

Para un alfabeto de 26 símbolos necesitamos ***log2(26) = 5 bits***

¿Es posible reducir el número de bits utilizados?

Compresión por probabilidad

Solución: Asignar códigos más cortos a los símbolos más frecuentes

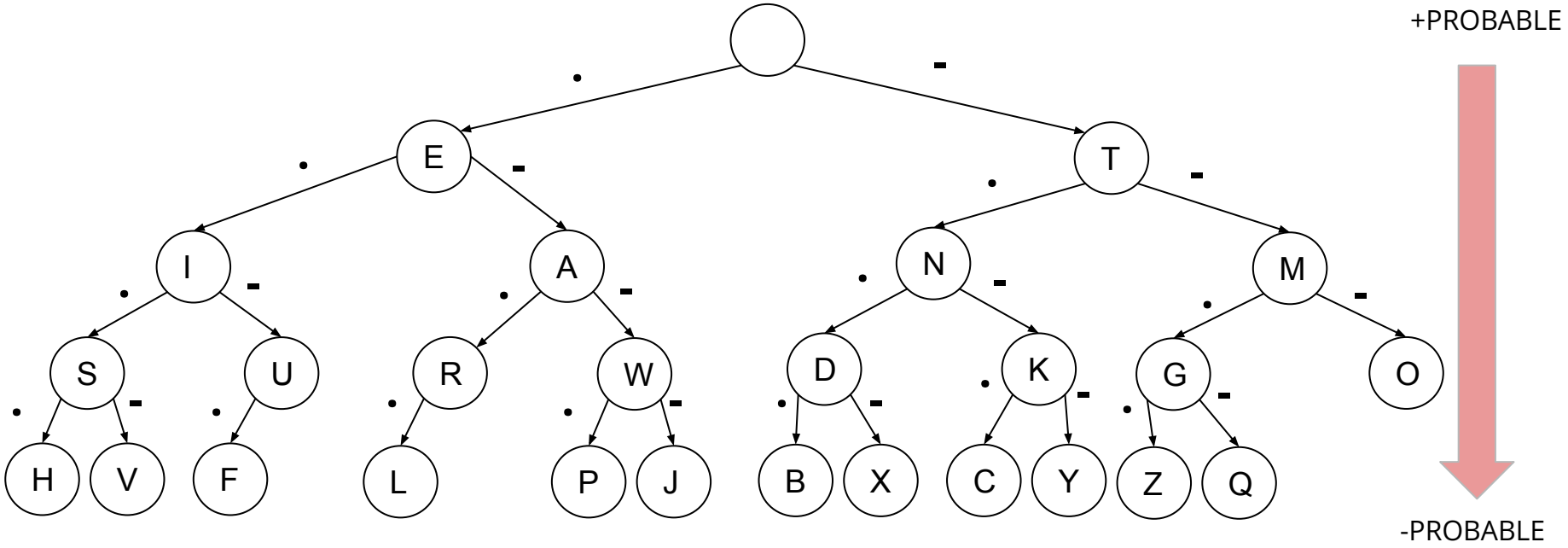


ASCII Alphabet			
A	1000001	N	1001110
B	1000010	O	1001111
C	1000011	P	1010000
D	1000100	Q	1010001
E	1000101	R	1010010
F	1000110	S	1010011
G	1000111	T	1010100
H	1001000	U	1010101
I	1001001	V	1010110
J	1001010	W	1010111
K	1001011	X	1011000
L	1001100	Y	1011001
M	1001101	Z	1011010

Part of the ASCII alphabet

Compresión por probabilidad

Ejemplo: Codificación de alfabeto inglés para código morse

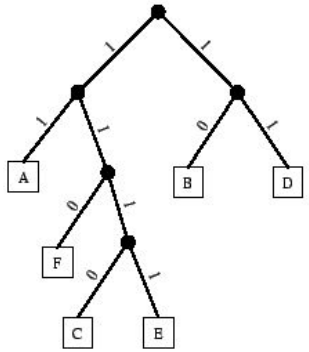


Compresión por probabilidad

Como el código morse como ejemplo, el resultado es un árbol con información en los nodos intermedios, lo cual se presta a ciertas ambigüedades.

La condición que debe cumplir una codificación para no presentar ambigüedades, es que la codificación de ningún carácter sea prefijo de otra.

Como por ejemplo:



Compresión por probabilidad

Tomemos el siguiente ejemplo:

Mensaje: "en mi casa viven tengo tres gatos"

letras del alfabeto: **e,n,m,i,c,a,s,v,t,g,o,r**

total de letras: **12**

Letra	repeticiones	probabilidad	codificación	cod. min	largo
e	4	4/27= **0.148**	0000	00	2
n	3	3/27= **0.111**	0010	10	2
m	1	1/27= **0.037**	1011	1011	4
i	2	2/27= **0.074**	0110	110	3
c	1	1/27= **0.037**	1010	1010	4
a	3	3/27= **0.111**	0100	100	3
s	3	3/27= **0.111**	0001	01	2
v	2	2/27= **0.074**	0101	101	3
t	3	3/27= **0.111**	0011	11	2
g	2	2/27= **0.074**	0111	111	3
o	2	2/27= **0.074**	1000	1000	4
r	1	1/27= **0.037**	1001	1001	4

log2((12) = 3.58 = 4 bits

Definición de Entropía y Redundancia

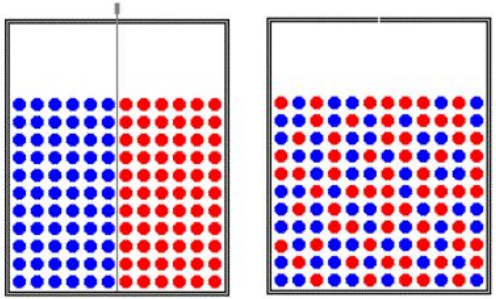
Redundancia:

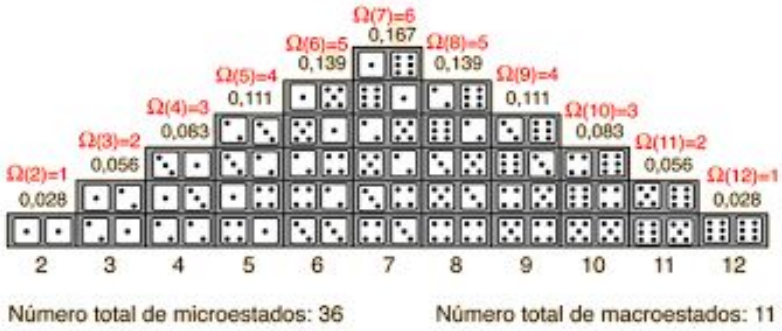
Datos repetitivos y predecibles

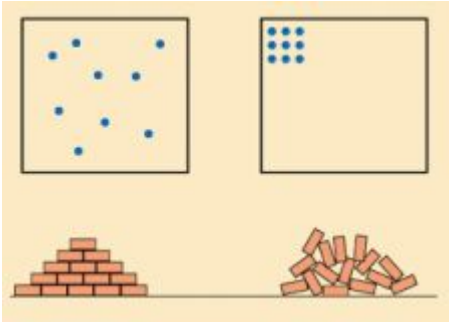
Entropía:

Mide la incertidumbre de una fuente de información, se puede considerar como la cantidad de información promedio que contienen los símbolos usados.

Definición de Entropía y Redundancia







Suma de los dos dados	Frecuencia	p	$\log_2(1/p)$	$p*\log_2(1/p)$
2	1	0.03	5.17	0.14
3	2	0.06	4.17	0.23
4	3	0.08	3.58	0.30
5	4	0.11	3.17	0.35
6	5	0.14	2.85	0.40
7	6	0.17	2.58	0.43
8	5	0.14	2.85	0.40
9	4	0.11	3.17	0.35
10	3	0.08	3.58	0.30
11	2	0.06	4.17	0.23
12	1	0.03	5.17	0.14
			H	3.27

Características de la entropía de datos

Los símbolos con menor probabilidad son los que aportan mayor información.

En la teoría de la información es una magnitud que mide la información provista por una fuente de datos.

La entropía es la encargada de medir la aleatoriedad en los datos.

Entropía de Shannon

Fue presentada por Shannon en su artículo de 1948, A Mathematical Theory of Communication.

Según Shannon la entropía debe satisfacer lo siguiente:

- La medida de información debe ser proporcional.
- Si todos los elementos de la señal son igual de probables a la hora de aparecer, entonces la entropía será máxima.

Entropía de Shannon

Shannon define la entropía del alfabeto como:

$$\text{Entropía} = -\sum P_i * \log2((P_i)$$

El teorema de Shannon dice que el número promedio de bits esperable para un conjunto de letras y probabilidades dadas se aproxima a la entropía del alfabeto.

Entropía de Shannon

Entropía = $-[(0.111 \log_2(0.111))+(0.037 \log_2(0.037))+(0.148 \log_2(0.148))+(0.074 \log_2(0.074))+(0.074 \log_2(0.074))+(0.037 \log_2(0.037))+(0.111 \log_2(0.111))+(0.111 \log_2(0.111))+(0.074 \log_2(0.074))]$

Entropía = $-[(-0.352)+(-0.176)+(-0.408)+(-0.278)+(-0.278)+(-0.176)+(-0.352)+(-0.278)+(-0.176)+(-0.352)+(-0.352)+(-0.278)]$

Entropia = $-[-3.45787]$

Entropia = **3.45787 bits/símbolo**