CÓDIGOS Y CRIPTOGRAFÍA. INFORMÁTICA. DIEGO RUANO 19/10 2022

SEGUNDA PARTE

CANALES SIN RUIDO. CÓDIGOS COMPRESORES

SI EL CANAL POR EL QUE CIRCULA LA INFORMACIÓN NO ES ALTERADO POR RUIDO, NUESTRO PROPÓSITO EN LA CODIFICACIÓN SERÁ REDUCIR LOS MENSAJES A SU FORMA MÁS CONCISA POSIBLE

¿CODIFICACIÓN SIN COMPRIMIR DE $\mathcal{A} = \{a, b, \ldots, x, y, z\}$?

UN VECTOR DE BITS LONGITUD 5 $(2^4 = 16 \leq 27 \leq 2^5 = 32)$

¿CÓDIGOS PARA COMPRIMIR?

↳ DE LONGITUD VARIABLE CON DESCODIFICACIÓN ÚNICA

SIMBOLOS MÁS FRECUENTES ⟶ MENOR LONGITUD

SIMBOLOS MENOS FRECUENTES ⟶ MAYOR LONGITUD

DESCODIFICACIÓN ÚNICA

PARA QUE EL CÓDIGO SEA ÚTIL DEBE TENER DESCODIFICACIÓN ÚNICA. PROBLEMA: LONGITUD VARIABLE ¿DONDE EMPIEZA Y TERMINA CADA PALABRA?

EJ: SI CODIFICAMOS $A=\{a,b,c\}$ ASÍ

FUENTE	CODIFICACIÓN
a	0
b	1
c	01

(a ES PREFIJO DE c, NO ES INSTANTANEO)

¿COMO DECODIFICAMOS 0101?

cc ó cab ó abc ó abab

NO ES ÚNICA → NO NOS VALE

SOLUCIONES:

- USAR UN SÍMBOLO SEPARADOR (COMO EN CÓDIGO MORSE) PERO NO ES UNA BUENA IDEA. SE PIERDE EFICIENCIA
- USAR UN CÓDIGO INSTANTANEO (PREFIX CODE)

DEF: DECIMOS QUE UN CÓDIGO ES INSTANTANEO SI NINGUNA PALABRA CÓDIGO ES PREFIJO DE LA OTRA

EJ: SI EL ALFABETO FUENTE ES $A = \{a, b, c\}$ Y LA FRECUENCIA CON QUE APARECE a ES MAYOR QUE b y c. PODEMOS CODIFICAR A COMO

FUENTE	COD I	COD II
a	0	0
b	01	10
c	11	11

NO INSTANTANEA AUNQUE TIENE DECODIFICACIÓN ÚNICA

INSTANTANEA

EJ : 0 1 1 1 1 . . . . . .

COD I : DEPENDE DEL FINAL $<$ a c . . . . . . c / b c c

$\underbrace{0}_{}\underbrace{11}_{}\underbrace{11}_{}\underbrace{11}_{}$ a c c c

$\underbrace{01}_{}\underbrace{11}_{}\underbrace{11}_{}$ a c c

Y TENEMOS QUE ESPERAR

NO ES ÚTIL ESPERAR TANTO

COD II : $\underbrace{0}_{a}\underbrace{1\,1}_{c}\underbrace{1\,1}_{c}$ . . . . .

Th : TODO CÓDIGO INSTANTANEO TIENE DESCODIFICACIÓN ÚNICA.

$\not\Leftarrow$ NO, ES CIERTO ES COD I, TIENE DESCODIFICACIÓN ÚNICA PERO NO ES INSTANTANEO

LONGITUD MEDIA DE UN CÓDIGO

$\mathcal{F}$ = FUENTE ASOCIADA AL ALFABETO $A = \{a_1, \dots, a_m\}$

CON DISTRIBUCIONES DE PROBABILIDAD $p_1, \dots, p_m$

$p_1 + \dots + p_m = 1$

SEA C UN CÓDIGO q-ARIO DE $\mathcal{F}$ DONDE LAS LONGITUDES DE LAS PALABRAS DE C SON $l_1, \dots, l_m$

(a_i APARECE CON PROBABILIDAD p_i Y SE CODIFICA EN UNA PALABRA DE LONGITUD l_i)

DEF: LONGITUD MEDIA DE C

$$\ell(C) = \sum_{i=1}^{m} p_i l_i$$

EJ: MENSAJE DE n SIMBOLOS SE CODIFICA EN $n \cdot \ell(C)$ SIMBOLOS DE MEDIA

TEOREMAS SHANNON

Th

$$\ell(C) \geq \frac{H(\mathcal{F})}{\log q}$$

FUNCIÓN ENTROPÍA:

$$H(\mathcal{F}) = \sum_{i=1}^{m} p_i \log\left(\frac{1}{p_i}\right)$$

DEF $\ell_q(\mathcal{F})$ MENOR LONGITUD MEDIA POSIBLE DE UN CÓDIGO INSTANTÁNEO q-ARIO DE LA FUENTE $\mathcal{F}$

Th

$$\frac{H(\mathcal{F})}{\log q} \leq \ell_q(\mathcal{F}) \leq \frac{H(\mathcal{F})}{\log q} + 1$$

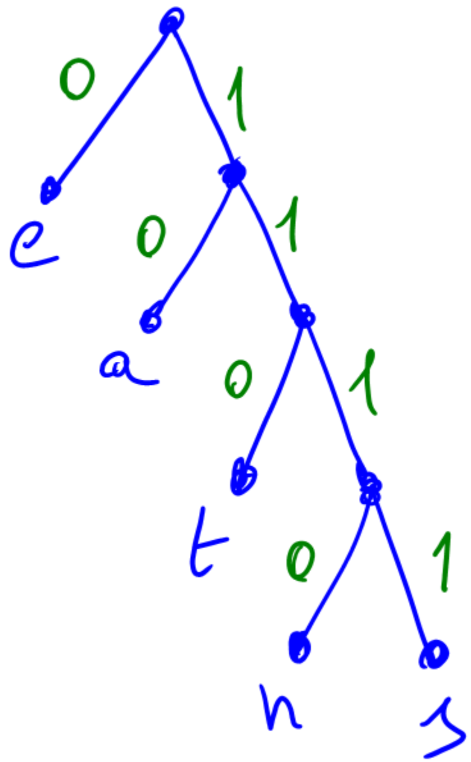
$$\log = \log_2$$

UNA CONSTRUCCIÓN ÓPTIMA: CÓDIGO HUFFMAN

VAMOS A REPRESENTAR EL CÓDIGO USANDO ÁRBOLES (GRAFOS)

$A = \{e, a, t, n, s\}$

EJ:



$e \longmapsto 0$

$a \longmapsto 10$

$t \longmapsto 110$

$n \longmapsto 1110$

$s \longmapsto 1111$

EJ: 1 1 1 1 1 0 1 1 1 0 0

CÓDIGO INSTANTANEO

INPUT ALGORITMO: FRECUENCIA DE CADA SIMBOLO

OUTPUT: CODIGO COMPRESOR ÓPTIMO

1) UN BOSQUE DE ÁRBOLES, DONDE CADA ÁRBOL TIENE UN ÚNICO VERTICE (UNO POR CADA SIMBOLO A CODIFICAR)
CADA VÉRTICE VIENE ETIQUETADO POR EL SIMBOLO Y CADA ÁRBOL POR SU FRECUENCIA.

2) ITERATIVAMENTE, COMBINAMOS LOS DOS ÁRBOLES CON MENOR FRECUENCIA, INTRODUCIENDO UNA NUEVA RAÍZ (PESO MAYOR → IZQUIERDA
PESO MENOR → DERECHA)

ETIQUETAMOS ARISTA IZQDA CON 0 Y LA DERECHA CON 1

ETIQUETAMOS EL NUEVO ÁRBOL CON LA SUMA DE LAS FRECUENCIAS.

EN CASO DE EMPATE ESCOGEMOS NOSOTROS

3) EL ALGORITMO SE TERMINA CUANDO YA NO TENEMOS UN BOSQUE, SI NO UN ÁRBOL

EJ: $A = \{A, B, C, D, E, F\}$

$P = \{0{,}08;\ 0{,}1;\ 0{,}12;\ 0{,}15;\ 0{,}2;\ 0{,}35\}$

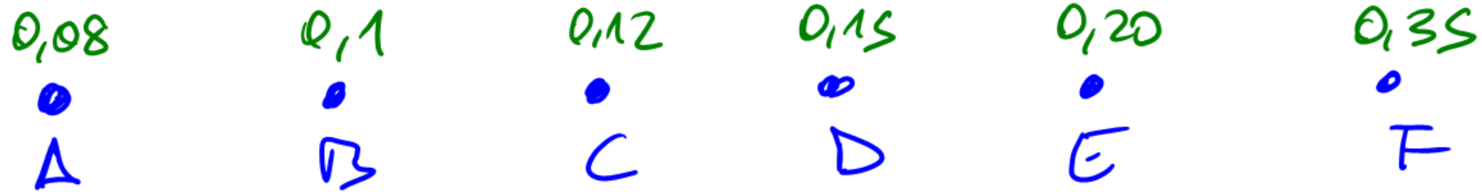
- NÚMERO MEDIO BITS SIN COMPRESIÓN

3 $\quad (4 = 2^2 \leq 6 \leq 2^3 = 8)$

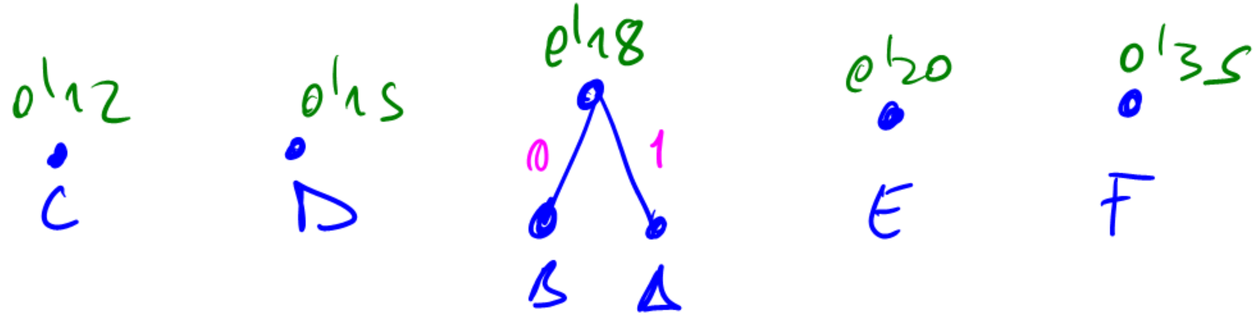
A 000
B 001 ...
C 010

- CON CÓDIGO HUFFMAN

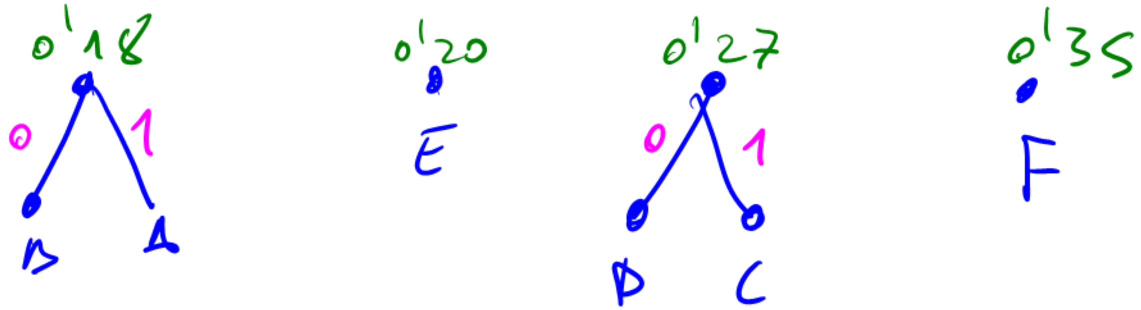
VAMOS A VERLO



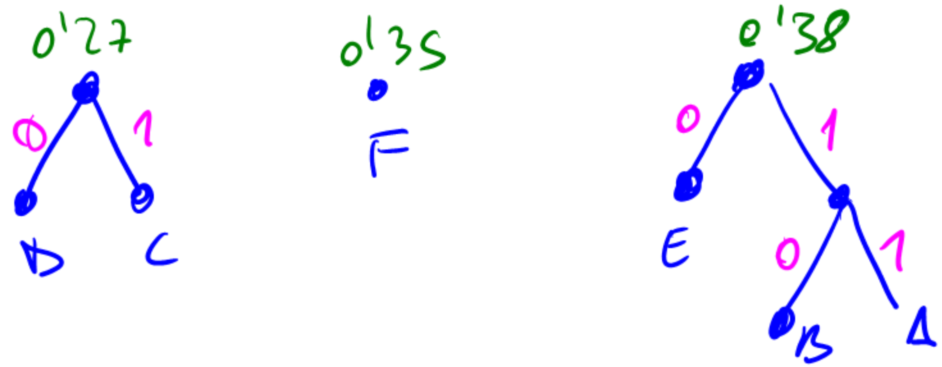
PASO 1



PASO 2

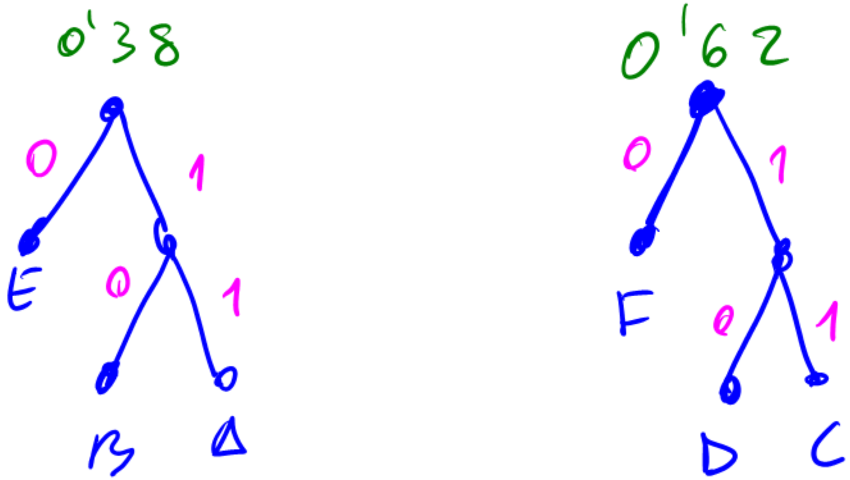


PASO 3



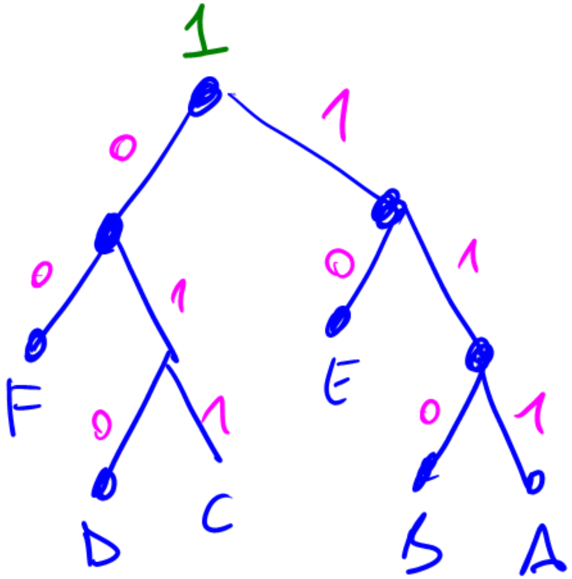
PASO 4

PASO 5



PASO 6

FINAL



A → 111

B → 110

C → 011

D → 010

E → 10

F → 00

NÚMERO MEDIO DE BITS USADOS CON HUFFMAN:

$$\mathcal{A} = \{A, B, C, D, E, F\}$$

$$\mathcal{P} = \{0{,}08, 0.1, 0{,}12, 0{,}15, 0{,}2, 0{,}35\}$$

$$l(C) = \sum_{i=1}^{m} p_i l_i$$

$$0'08 \cdot 3 + 0'1 \cdot 3 + 0'12 \cdot 3 + 0'15 \cdot 3 + 0'2 \cdot 2 + 0'35 \cdot 2 =$$

$$= 2'45 < 3$$

TEXTO DE 100 SÍMBOLOS. CON HUFFMAN VAMOS A USAR DE MEDIA 245 BITS

SIN COMPRESIÓN USAMOS 300 BITS

Th: EL CÓDIGO DE HUFFMAN ES ÓPTIMO (PARA CÓDIGOS BINARIOS)

SE PUEDE EXTENDER LA CONSTRUCCIÓN PARA CÓDIGOS SOBRE $\mathbb{F}_q$