Aula 04


Universidade Federal de Santa Maria
SEDES SAPIENTIAE
1960

Compressão por entropia

- Shannon
- Shannon-Fano
- Huffman

Compressão por entropia

Mensagem:
“o doce perguntou pro doce qual é o doce mais doce que o doce de batata doce o doce respondeu pro doce que o doce mais doce que o doce de batata doce é o doce de doce de batata doce”

Total: 180 caracteres

Mensagem = 180 * 8bits = 1440 bits

Caracter	Frequência	Probabilidade	Entropia
' '	40	0,22	-0,48
'o'	26	0,14	-0,40
'e'	25	0,14	-0,40
'd'	20	0,11	-0,35
'C'	15	0,08	-0,30
'a'	12	0,07	-0,26
'u'	7	0,04	-0,18
't'	7	0,04	-0,18
'q'	4	0,02	-0,12
'p'	4	0,02	-0,12
'r'	4	0,02	-0,12
'b'	3	0,02	-0,10
's'	3	0,02	-0,10
'é'	2	0,01	-0,07
'i'	2	0,01	-0,07
'm'	2	0,01	-0,07
'n'	2	0,01	-0,07
'g'	1	0,01	-0,04
'l'	1	0,01	-0,04
	180	1	**3,49**

Compressão por entropia

Shannon-Fano

Shannon

Fano

Algoritmo

- Decidir o tamanho de código para cada caractere
- Escolher um código do tamanho decidido

Tamanho de códigos

$$[-\log_2 p_i]$$

Código

Manualmente (escolher primeiro código - lexicograficamente - que mantenha as propriedades de *prefix-free*)

Usar probabilidades cumulativas

Compressão por entropia

Shannon-Fano

Shannon

Fano

Caracter	Shannon (bits)	$(-\log P(x))_2$ c.	Código
' '	3	0.00000000	000
'o'	3	0.00111000	001
'e'	3	0.01011101	010
'd'	4	0.10000001	1000
'c'	4	0.10011101	1001
'a'	4	0.10110011	1011
'u'	5	0.11000100	11000
't'	5	0.11001110	11001
'q'	6	0.11011000	110110
'p'	6	0.11011101	110111
'r'	6	0.11100011	111000
'b'	6	0.11101001	111010
's'	6	0.11101101	111011
'é'	7	0.11110001	1111000
'i'	7	0.11110100	1111010
'm'	7	0.11110111	1111011
'n'	7	0.11111010	1111101
'g'	8	0.11111101	11111101
'l'	8	0.11111110	11111110

Universidade Federal de Santa Maria
1960

Compressão por entropia

Shannon-Fano

Shannon

Fano

Total = (40 * 3b) + (26 * 3b) + (25 * 3b) + (20 * 4b) + (15 * 4b) + (12 * 4b) + (7 * 5b) + (7 * 5b) + (4 * 6b) + (4 * 6b) + (4 * 6b) + (3 * 6b) + (3 * 6b) + (2 * 7b) + (2 * 7b) + (2 * 7b) + (2 * 7b) + (1 * 8b) + (1 * 8b)

Total = 711 bits

T_c = 711b / 1440b = 0,49375

bits/caracter = 711b/180c = **3,9889 bits/caracter**

Compressão por entropia

Shannon-Fano

Shannon

Fano

Algoritmo

1 Ordenar caracteres por probabilidade (descendente)

2 Dividir o conjunto em 2 grupos com probabilidades mais próximas possível da igualdade

3 Repetir passo 2 para cada grupo de caracteres até não ser possível dividir mais, montando uma árvore

4 Cada sub-árvore esquerda recebe o prefixo 0 e direita 1

Universidade Federal de Santa Maria
SEDES SAPIENTIÆ
1960

Compressão por entropia

Shannon-Fano

Shannon

Fano

Caracter	Frequência	Probabilidade	Entropia
' '	40	0,22	-0,48
'o'	26	0,14	-0,40
'e'	25	0,14	-0,40
'd'	20	0,11	-0,35
'C'	15	0,08	-0,30
'a'	12	0,07	-0,26
'u'	7	0,04	-0,18
't'	7	0,04	-0,18
'q'	4	0,02	-0,12
'p'	4	0,02	-0,12
'r'	4	0,02	-0,12
'b'	3	0,02	-0,10
's'	3	0,02	-0,10
'é'	2	0,01	-0,07
'i'	2	0,01	-0,07
'm'	2	0,01	-0,07
'n'	2	0,01	-0,07
'g'	1	0,01	-0,04
'l'	1	0,01	-0,04
	180	1	**3,49**

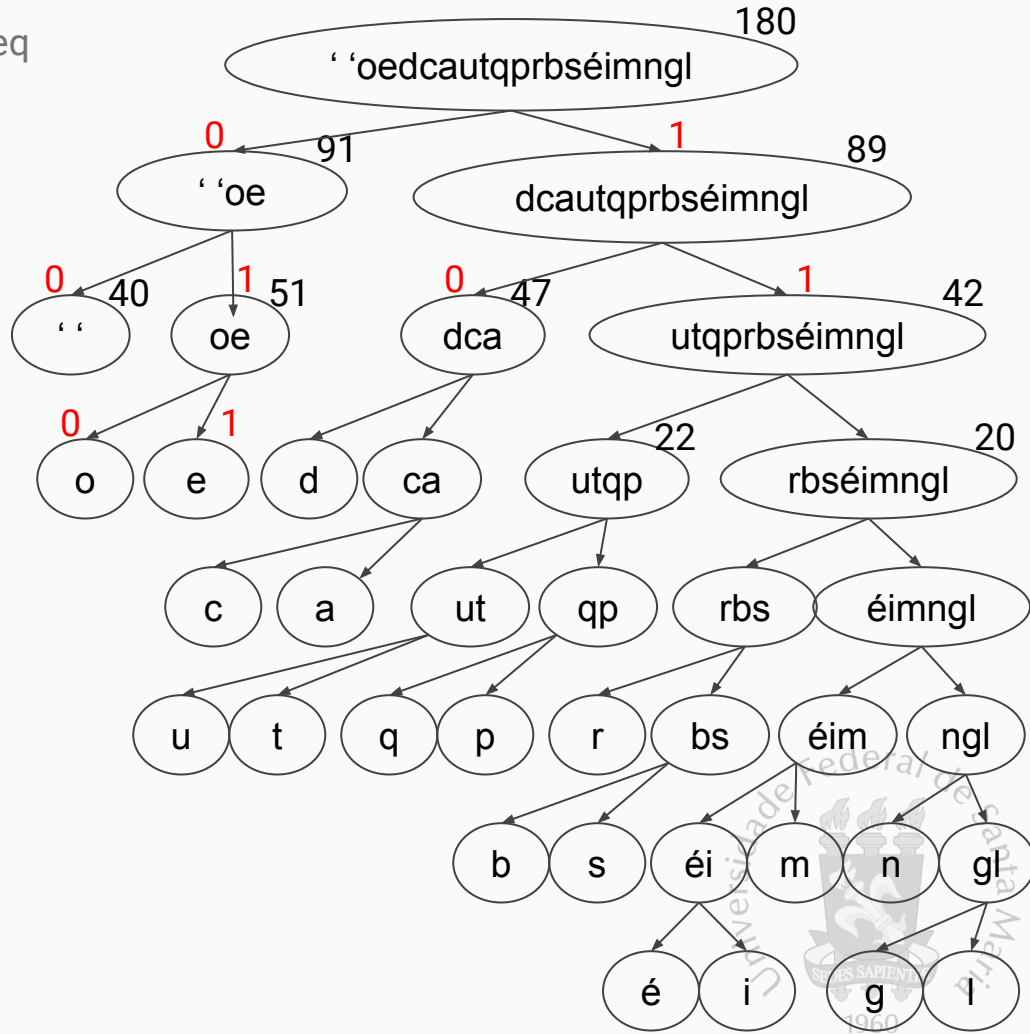
Compressão por entropia

Shannon-Fano

Shannon

Fano

Car.	Freq
‘ ’	40
'o'	26
'e'	25
'd'	20
'c'	15
'a'	12
'u'	7
't'	7
'q'	4
'p'	4
'r'	4
'b'	3
's'	3
'é'	2
'i'	2
'm'	2
'n'	2
'g'	1
'l'	1



Compressão por entropia

Shannon-Fano

Shannon

Fano

Car.	Freq	Código	bits
' '	40	00	2
'o'	26	010	3
'e'	25	011	3
'd'	20	100	3
'c'	15	1010	4
'a'	12	1011	4
'u'	7	11000	5
't'	7	11001	5
'q'	4	11010	5
'p'	4	11011	5
'r'	4	11100	5
'b'	3	111010	6
's'	3	111011	6
'é'	2	1111000	7
'i'	2	1111001	7
'm'	2	111101	6
'n'	2	111110	6
'g'	1	1111110	7
'l'	1	1111111	7
	180		Tot=633bits 633b/1440b=0,43958

713b/180c = **3,5167 bits/caracter**

Universidade Federal de Santa Maria
1960