Aula 05


Universidade Federal de Santa Maria
SEDES SAPIENTIAE
1960

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

Caracter	Shannon (bits)	$(-\log P(x))_2$	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

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

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

633b/180c = **3,5167 bits/caracter**

Problemas com Shannon-Fano

Tabela de conversão não é fixa

Logo

Deve ser enviada junto à mensagem codificada

Nenhum destes métodos garante código ótimo

Codificação de Huffman

Método ótimo

Implementação simples (e de tempo linear → O(n))

Ótimo para codificação de caracteres isolados

Há métodos com melhores taxas de compressão

Frequentemente usado como parte de um processo de compressão mais elaborado (ex.: MP3, JPEG, etc)

Compressão por entropia

Huffman

Algoritmo

1) Criar uma fila de caracteres por probabilidade, decrescente. Considerar que todos são folhas, ainda desconexas.
2) Retirar os dois com menor probabilidade, criando um novo nó. Inserir novo nó na fila, considerando que seus componentes são seus filhos.
3) Enquanto houver mais de um nó na fila, repetir passo 2.
4) Nó remanescente na fila é a raiz da árvore.

A montagem da tabela de codificação/decodificação é similar à de Fano: filhos à esquerda recebem prefixo 0 e à direita prefixo 1.

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”

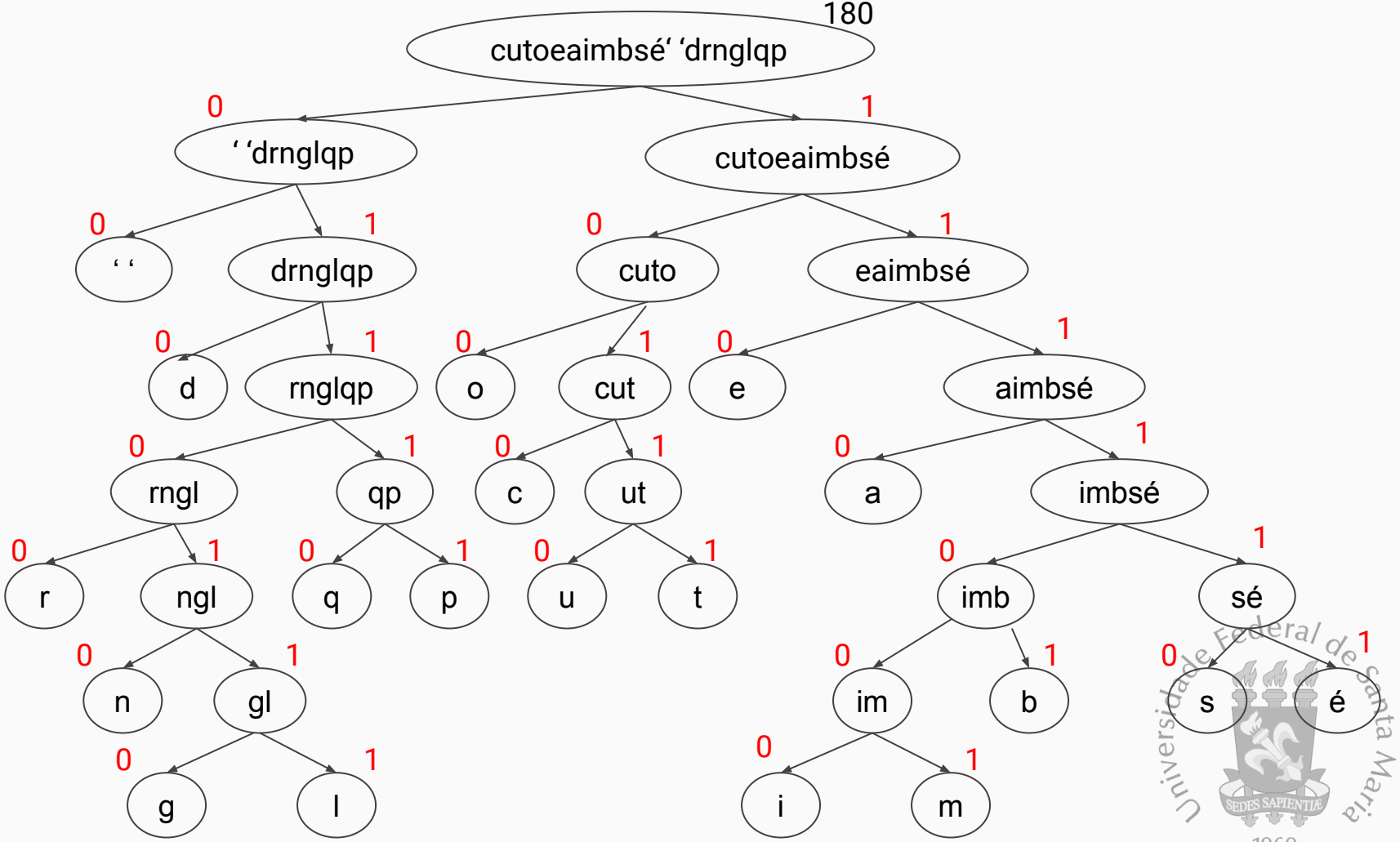
Total: 180 caracteres

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**

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

1) (‘ ‘,o,e,d,c,a,u,t,q,p,r,b,s,é,i,m,n,g,l)
(40,26,25,20,15,12,7,7,4,4,4,3,3,2,2,2,2,1,1)
2) (‘ ‘,o,e,d,c,a,u,t,q,p,r,b,s,é,i,m,n,gl)
(40,26,25,20,15,12,7,7,4,4,4,3,3,2,2,2,2,2)
3) (‘ ‘,o,e,d,c,a,u,t,q,p,r,ngl,b,s,é,i,m)
(40,26,25,20,15,12,7,7,4,4,4,4,3,3,2,2,2)
4) (‘ ‘,o,e,d,c,a,u,t,q,p,r,ngl,im,b,s,é)
(40,26,25,20,15,12,7,7,4,4,4,4,4,3,3,2)
5) (‘ ‘,o,e,d,c,a,u,t,sé,q,p,r,ngl,im,b)
(40,26,25,20,15,12,7,7,5,4,4,4,4,4,3)
6) (‘ ‘,o,e,d,c,a,u,t,imb,sé,q,p,r,ngl)
(40,26,25,20,15,12,7,7,7,5,4,4,4,4)
7) (‘ ‘,o,e,d,c,a,rngl,u,t,imb,sé,q,p)
(40,26,25,20,15,12,8,7,7,7,5,4,4)
8) (‘ ‘,o,e,d,c,a,rngl,qp,u,t,imb,sé)
(40,26,25,20,15,12,8,8,7,7,7,5)
9) (‘ ‘,o,e,d,c,a,imbsé,rngl,qp,u,t)
(40,26,25,20,15,12,12,8,8,7,7)
10) (‘ ‘,o,e,d,c,ut,a,imbsé,rngl,qp)
(40,26,25,20,15,14,12,12,8,8)
11) (‘ ‘,o,e,d,rnglqp,c,ut,a,imbsé)
(40,26,25,20,16,15,14,12,12)
12) (‘ ‘,o,e,aimbsé,d,rnglqp,c,ut)
(40,26,25,24,20,16,15,14)
13) (‘ ‘,cut,o,e,aimbsé,d,rnglqp)
(40,29,26,24,25,20,16)
14) (‘ ‘,drnglqp,cut,o,e,aimbsé)
(40,36,29,26,24,25)
15) (eaimbsé,‘ ‘,drnglqp,cut,o)
(49,40,36,29,26)
16) (cuto,eaimbsé,‘ ‘,drnglqp)
(55,49,40,36)
17) (‘ ‘drnglqp,cuto,eaimbsé)
(76,55,49)
18) (cutoeaimbsé,‘ ‘drnglqp)
(104,76)
19) (cutoeaimbsé‘ ‘drnglqp)
(180)



Compressão por entropia

Huffman

Car.	Freq	Código	bits
' '	40	00	2
'o'	26	100	3
'e'	25	110	3
'd'	20	010	3
'c'	15	1010	4
'a'	12	1110	4
'u'	7	10110	5
't'	7	10111	5
'q'	4	01110	5
'p'	4	01111	5
'r'	4	01100	5
'b'	3	111101	6
's'	3	111110	6
'é'	2	111111	6
'i'	2	1111000	7
'm'	2	1111001	7
'n'	2	011010	6
'g'	1	0110110	7
'l'	1	0110111	7
	180		Tot=633bits 633b/1440b=0,43958

633b/180c = **3,5167 bits/caracter**

Comparação

Shannon

Fano

Huffman

caracter	A	B	C	D	E
frequência	15	7	6	6	5
probabil.	0.385	0.179	0.154	0.154	0.128
Shannon (código)	00	011	100	101	110
Fano (código)	00	01	10	110	111
Huffman (código)	1	000	001	010	011

codificação	Shannon	Fano	Huffman
bits/caracter	≈ 2,62	≈ 2,28	≈ 2,23