SD Curs 9

30 Apr 2025

1. Coduri Huffman

Coduri Huffman

Text

frecvența literelor

a	b	c	d	e	f
45	20	15	13	8	6

Problemă: Să se găsească o codificare în cod binar pentru fiecare literă astfel încât lungimea totală a textului codificat să fie minimă. De asemenea, vrem să putem decodifica textul binar.

a = 000
b = 001
c = 001 ⟿ lungimea totală a textului
d = 010 $107 \times 3 = 321$
e = 011
f = 101

Exemplu de codificare: a a b d = 000 000 001 011

010 101 011 = c f d

Notăm alfabetul cu $\mathcal{C}$ și pentru fiecare literă $c \in \mathcal{C}$ notăm $f(c)$ frecvența, dorim să minimizăm

$$\sum_{c \in \mathcal{C}} f(c) \text{ lungimea codului lui } c$$

Exemplu ambiguitate

a	b	c
0	01	10

? $\overset{a}{0}\ \overset{b}{\overline{0\ 1}}\ \overset{a}{0}$

$\underset{a}{0}\ \underset{a}{0}\ \underset{c}{1\ 0}$

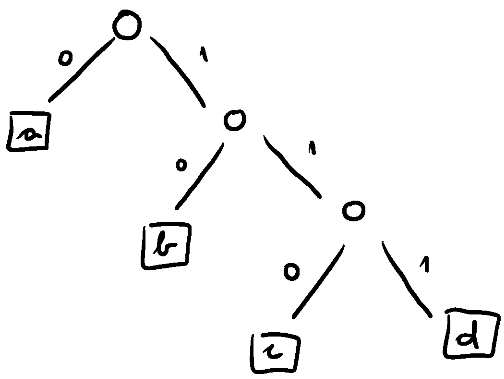
Pentru a putea decodifica, ne focusăm pe coduri care nu sunt prefixe ale altor coduri

Exemplu de cod „prefix-free"

a	b	c	d
0	10	110	111

$\underset{a}{0}\ \underset{b}{10}\ \underset{c}{110}\ \underset{b}{10}\ \underset{d}{111}$

Orice cod „prefix-free" poate fi reprezentat ca un arbore binar.



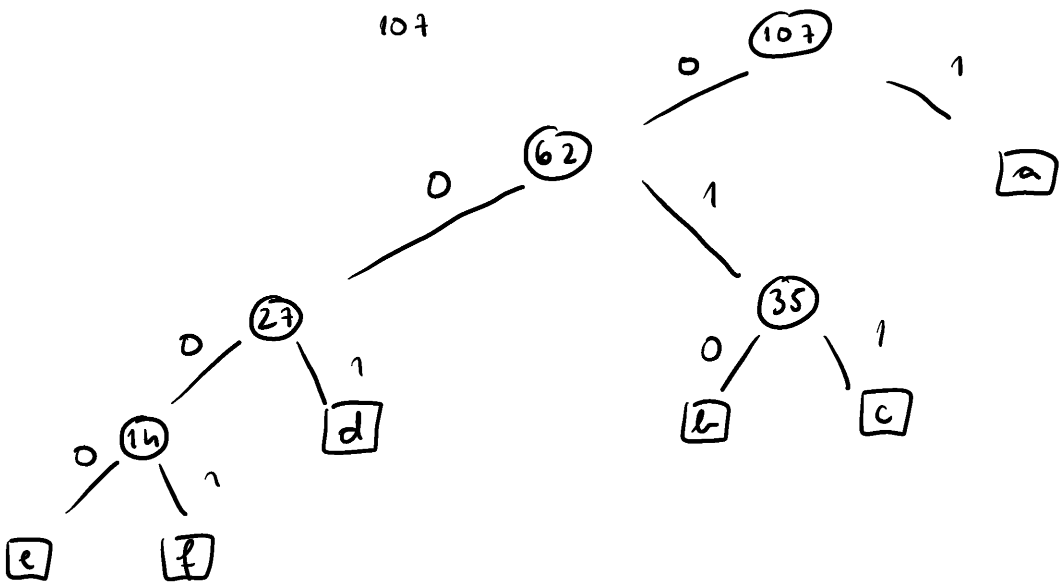
$$\min \sum_{c \in \mathcal{C}} f(c) \cdot d(c)$$

adâncimea frunzei corespunzătoare în arborele T

Algoritmul care găsește codul optim

- luăm 2 litere ce au frecvența minimă. Le punem în arbore ca fii ai unui nou nod care are frecvența suma celor două frecvențe

a	b	c	d	e	f
45	20	15	13	$\cancel{8}$	$\cancel{6}$
45	20	15	$\cancel{13}$	$\cancel{14}$	
45	$\cancel{20}$	$\cancel{15}$	27		
45	$\cancel{35}$	$\cancel{27}$			
$\cancel{45}$	$\cancel{62}$				
107					



a	b	c	d	e	f
1	101	011	001	0000	0001

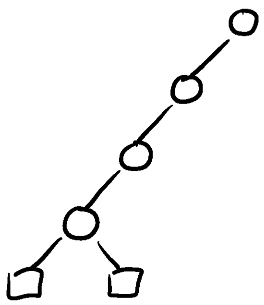
$$45 \cdot 1 + 20 \cdot 3 + 15 \cdot 3 + 13 \cdot 3 + 8 \cdot 4 + 6 \cdot 4 = \boxed{245}$$

vs 321

<u>Obs</u> (exercițiu pt. acasă)

Un cod optim corespunde unui arbore binar plin (în care fiecare nod cu excepția frunzelor are exact 2 fii).

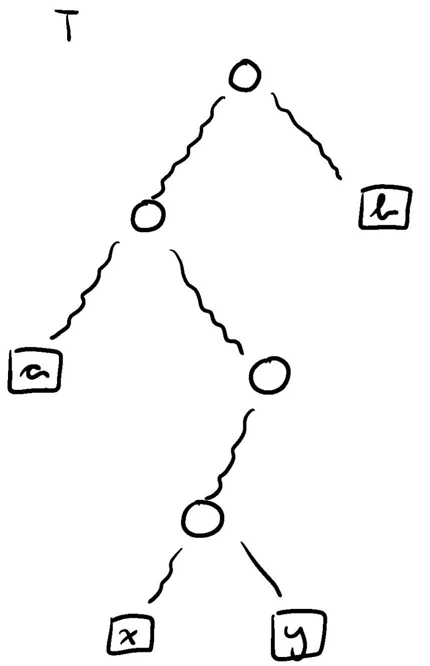
De ce nu așa?

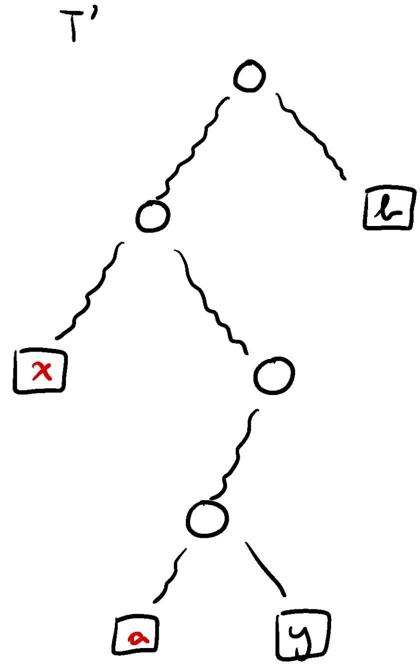


<u>Teorema 1</u> Fie $\mathcal{C}$ un set de caractere și T arborele corespunzător codului optim. Fie $a, b \in \mathcal{C}$ caracterele de frecvență minimă.

Atunci a și b vor fi frați în arbore și vor fi pe nivelul cel mai de jos.

<u>Dem</u> Pp. prin R.A.





nodurile cu adancime maximă

Construim un arbore T' în care interschimbăm nodurile a și x (unul din nodurile cu adancime max)

Presupunem $f(a) < f(b)$

depth

$$\text{cost}(T) = \sum_{c \in C} f(c) \cdot d(c)$$

Arătăm $\text{cost}(T') - \text{cost}(T) \leq 0$

$$\text{cost}(T') - \text{cost}(T) = f(a) \cdot d_{T'}(a) + f(x) \cdot d_{T'}(x) - f(a) \cdot d_T(a) - f(x) \cdot d_T(x)$$

$$= f(a)\left(\overset{d_T(x)}{\overset{\parallel}{d_{T'}(a)}} - d_T(a) \right) + f(x)\left(\underset{d_T(a)}{\underset{\parallel}{d_{T'}(x)}} - d_T(x) \right)$$

$$= \underset{\geq 0}{\left(d_T(x) - d_T(a) \right)} \underset{\leq 0}{\left(f(a) - f(x) \right)}$$

Repetăm procedura cu b și y și obținem arborele din teoremă

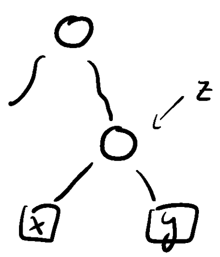
Teorema 2 Fie $\mathcal{C}$ un set de caractere și T arborele corespunzător codului optim. Fie $x, y \in \mathcal{C}$ caracterele de frecvență minimă.

Fie $\mathcal{C}' = \mathcal{C} - \{x, y\} \cup \{z\}$ unde $f(z) = f(x) + f(y)$

Fie T' arborele optim pt $\mathcal{C}'$. Atunci T obținut din T' în care înlocuim pe z cu un nod intern ce are ca fii pe x și y este arbore optim pentru $\mathcal{C}$

$\mathcal{C}'$ T'

$\mathcal{C}$ T



Dem Pp. prin R.A. că T nu e optim pt $\mathcal{C}$.
Deci există T'' cu $cost(T'') < cost(T)$

Arăt că există T''' cu
cu $cost(T''') < cost(T')$ un arbore cu cost mai mic pt $\mathcal{C}'$ (contradicție)

$$\text{cost}(T) - \text{cost}(T') = f(a) \cdot \underbrace{d_T(a)}_{=\, d_{T'}(z)+1} + f(b) \cdot \underbrace{d_T(b)}_{=\, d_{T'}(z)+1} - \underbrace{f(z)}_{=\, f(a)+f(b)} \cdot d_{T'}(z)$$

$$= f(a)\cancel{d_{T'}(z)} - f(a)\cancel{d_{T'}(z)} + f(b)\cancel{d_{T'}(z)} - f(b)\cancel{d_{T'}(z)} + f(a) + f(b)$$

Dacă $\exists$ T'' cu $\text{cost}(T'') < \text{cost}(T)$

înlocuind în T'' cele două noduri cu frecvență minimă cu nou nod ce are frecvență suma lor

$\Rightarrow$ T''' obținut astfel are $\text{cost}(T''') = \text{cost}(T'')^{*} - f(a) - f(b)$

$$\text{cost}(T''') < \text{cost}(T) - f(a) - f(b) = \text{cost}(T')$$