ENTROPIA

- **entropia**
  - valoarea medie de informației primită despre o variabilă X

$$H(X) = E(I(X)) = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} = \sum_{i=1}^{N} -p_i \log_2 p_i$$

  - *H(X)* se numește entropia lui X
  - *I(X)* este informația despre X
  - *E* este “expected value”, operația care calculează valoarea medie

  - exemplu: X = {A, B, C, D} cu probabilități {1/3, ½, 1/12, 1/12}
    - $H(X) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{12}\log_2\frac{1}{12} - \frac{1}{12}\log_2\frac{1}{12} = 1.626 \text{ biti}$
    - variabila X are 4 opțiuni, deci în mod normal am avea nevoie de 2 biți să memorăm toate posibilitățile, dar (pentru că probabilitățile nu sunt egale) putem să codăm mai bine de 2 biți

CODAREA DATELOR

- **cum atingem acel 1.626 biți pentru cele patru evenimente?**
- **folosim o codare diferită de cea standard**
  - codarea standard A = 00, B = 01, C = 10, D = 11
  - cu această codare avem ABBC = 00 01 01 10
  - decodarea este directă: luăm câte 2 biți și fiecare e un eveniment

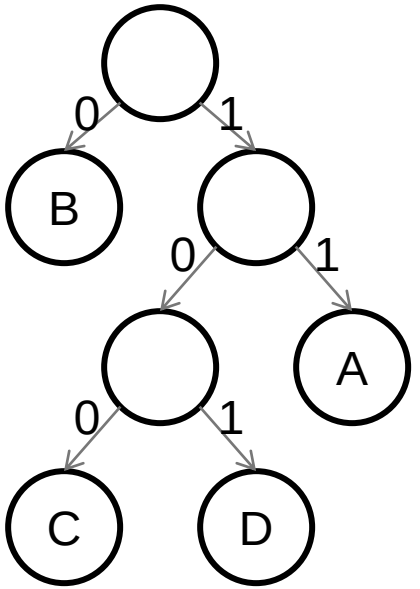
- **o altă codare (mai eficientă):**
  - dimensiune variabilă a codului: A = 01, B = 1, C = 000, D = 001
  - acum avem ABBC = 01 1 1 000
  - acum sunt 7 biți, față de 8 înainte (deci e mai bine)
  - decodarea trebuie să fie unică! trebuie să ne putem întoarce

- **o altă codare (și mai eficientă, dar incorectă):**
  - dimensiune variabilă a codului: A = 1, B = 0, C = 10, D = 11
  - acum avem ABBC = 1 0 0 10
  - 5 biți, și mai bine
  - problema? decodarea: dacă primim 10010 cum îl decodăm?
    - ABBC sau CBC sau …

CODAREA DATELOR

- **cum putem crea o codare eficientă și unică?**
  - un arbore binar
    - frunzele sunt codurile
    - stânga/dreapta e decis de 0/1
    - codarea este:
      - B = 0
      - A = 11
      - C = 100
      - D = 101
    - asta garantează codare eficientă

    și decodare unică



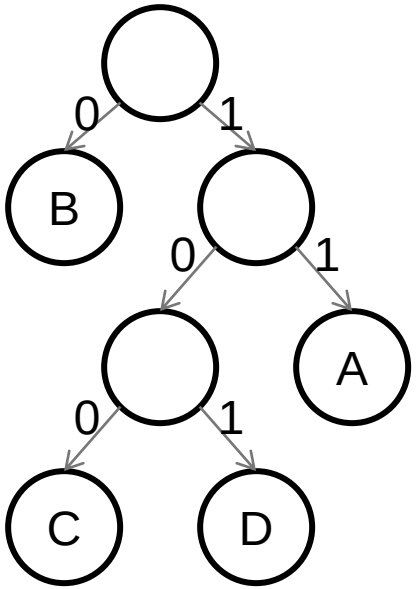
  - cum generăm codarea eficientă?
    - algoritmul Huffman
    - input: probabilitatea fiecărui eveniment {1/3, ½, 1/12, 1/12}
    - output: codurile care se citesc de pe un arbore binar (mai sus)
    - cheia: unele evenimente/simboluri apar mai des decât altele, deci acestea primesc o codare mai scurtă
    - dacă toate evenimente sunt equiprobabile, atunci nu putem face nimic

https://en.wikipedia.org/wiki/Huffman_coding

CODAREA DATELOR

- **cum putem crea o codare eficientă și unică?**
  - un arbore binar
    - frunzele sunt codurile
    - stânga/dreapta e decis de 0/1
    - codarea este:
      - B = 0
      - A = 11
      - C = 100
      - D = 101
    - asta garantează codare eficientă

    și decodare unică



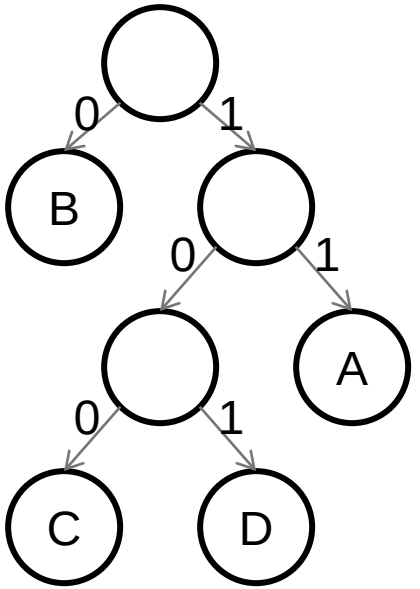
  - exercițiu: decodați 0 0 100 11 101 0 0 11
    - soluția: BBCADBBA

https://en.wikipedia.org/wiki/Huffman_coding

CODAREA DATELOR

- **cum putem crea o codare eficientă și unică?**
  - un arbore binar
    - frunzele sunt codurile
    - stânga/dreapta e decis de 0/1
    - codarea este:
      - B = 0
      - A = 11
      - C = 100
      - D = 101
    - asta garantează codare eficientă

    și decodare unică



- exercițiu: cum calculăm eficiența acestei codări? dimensiunea in medie a unui mesaj este? probabilitățile sunt {1/3, ½, 1/12, 1/12}
  - 2x1/3 + 1x1/2 + 3x1/12 + 3x1/12 = 1.667 biți
  - comparat cu 1.626 biți care e optim

https://en.wikipedia.org/wiki/Huffman_coding