Mã hóa – Mã thống kê tối ưu

- Khái niệm mã hóa, các thông số của mã hóa
- Mã thống kê
  - Entropy
  - Mã Shannon-Fano
  - Mã Huffman

Mã thống kê – Khái niệm về Entropy

- Entropy trong lí thuyết thông tin là phép đo định lượng về "thông tin" của nguồn tin.
  - Nguồn tin có Entropy lớn ⇔ nội dung ngẫu nhiên
  - Nguồn tin có Entropy nhỏ ⇔ nội dung có có cấu trúc, lặp lại.
- Entropy được sử dụng trong việc mã hóa – nén thông tin. Nếu phân bố xác suất PDF của nguồn tin được biết trước, giá trị Entropy cho biết số bit trung bình cần thiết để mã hóa nguồn tin.

Mã thống kê – Tính giá trị Entropy

$$H(X) = -\sum_{x \in X} p(x).\log_b\ p(x)$$

- *H(X)* – Entropy của nguồn tin
- *X* – Nguồn tin với các kí tự *x*
- *b=2* - bit thông tin

Ví dụ:

symbol	Tần suất	p(x)	-p(x).log2p(x)
a	5	0.45	0.52
b	2	0.18	0.45
r	2	0.18	0.45
c	1	0.09	0.31
d	1	0.09	0.31
	11		2.04

H(X)=2.04

Mã thống kê – Tính chất của Entropy

$$H(X) = -\sum_{x \in X} p(x).\log_b\ p(x)$$

Ví dụ: Nguồn tin "*abracadabra*"

symbol	Tần suất	p(x)	-p(x).log2p(x)
a	5	0.45	0.52
b	2	0.18	0.45
r	2	0.18	0.45
c	1	0.09	0.31
d	1	0.09	0.31
	11		2.04

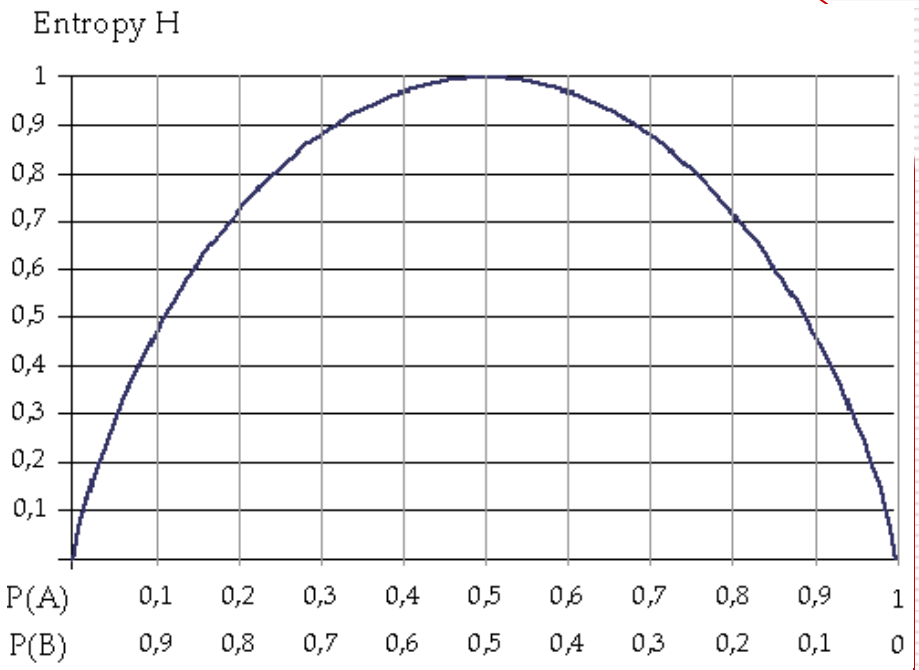
H(X)=2.04

Nguồn tin "*abracadabra*" có thể mã hóa với mã có độ dài trung bình 2.04bit/kí tự. Bản tin mã hóa theo cách này được gọi là **mã tối ưu** hay **mã hóa Entropy**.

Mã thống kê – Entropy của nguồn tin nhị phân

Bản tin binary gồm 2 kí tự A,B

P(A)=1-P(B)



Nhận xét:

- Giá trị Entropy cực đại H=1 khi A và B có xác suất như nhau (0.5). Khi đó độ dài mã trung bình là 1 bit – tối ưu.
- Trong các trường hợp còn lại, H<1, cần lựa chọn mã khác để đạt hiệu quả tốt hơn (code efficiency)

Mã thống kê – Định nghĩa và phân loại

- Entropy cung cấp thông tin về độ dài từ mã cần thiết cho việc mã hóa nguồn tin.
- Điều kiện tiên quyết của mã thống kê là cần biết trước xác suất xuất hiện của các kí tự (symbol) trong nguồn tin.
- Bộ mã hóa thống kê sẽ gán các từ mã (code word) có độ dài ngắn vào các kí tự có xác suất lớn, và ngược lại, gán từ mã có độ dài lớn cho các kí tự có xác suất nhỏ => Giảm kích thước của nguồn tin.
- Các thuật toán của mã hóa thống kê
  - Mã Shannon-Fano
  - Mã Huffman

Mã Shannon-Fano

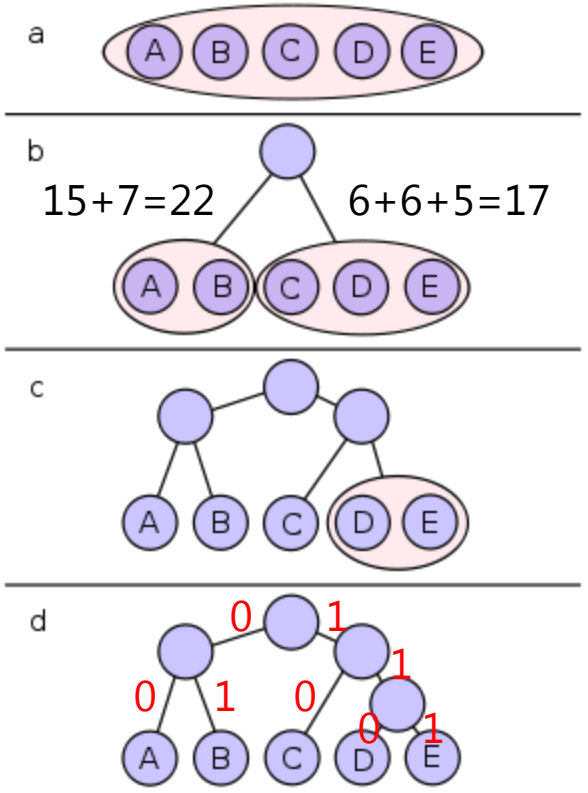
- Do Shannon và Fano độc lập xây dựng dựa trên lí thuyết Entropy.
- Mã Shannon-Fanon được xây dựng nhằm tối ưu hóa độ dài của từng từ mã (code word) tiệm cận với giá trị -log*p(x)*.

Ví dụ:

symbol	Tần suất	p(x)	**Lượng tin riêng** -log2p(x)
A	15	0.38	1.38
B	7	0.18	2.48
C	6	0.15	2.70
D	6	0.15	2.70
E	5	0.13	2.96

H(X)=2.1858

symbol	**Code word**
A	00
B	01
C	10
D	110
E	111



Mã Huffman

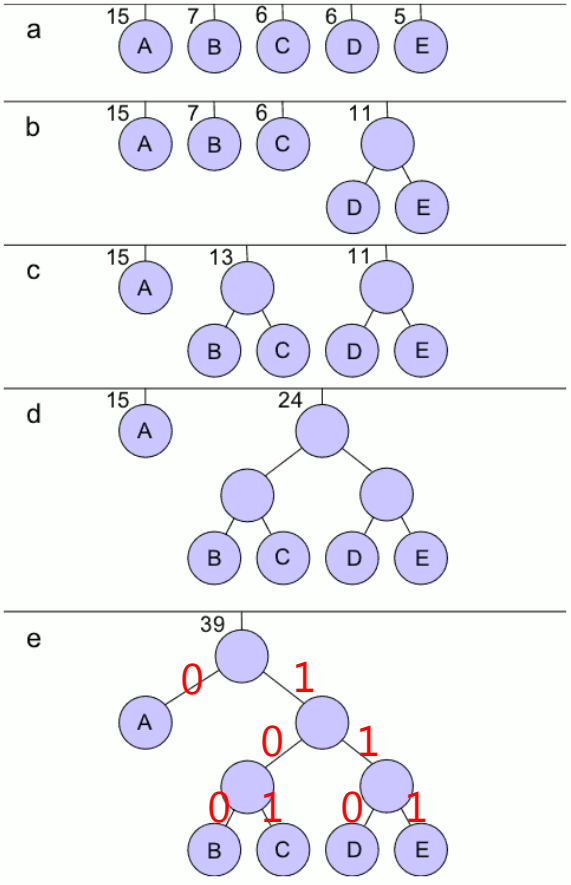
- Mã Huffman được xây dựng dựa trên lí thuyết Entropy
- Mã Huffman xây dựng cây nhị phân và gán giá trị bit từ dưới lên (bottom-up) nhằm tối ưu hóa kích thước của toàn bộ bản tin.

Ví dụ:

symbol	Tần suất	p(x)	**Lượng tin riêng** -log2p(x)
A	15	0.38	1.38
B	7	0.18	2.48
C	6	0.15	2.70
D	6	0.15	2.70
E	5	0.13	2.96

H(X)=2.1858

symbol	**Code word**
A	0
B	100
C	101
D	110
E	111



So sánh giữa mã Shannon-Fano và Huffman

- Mã Shannon-Fano: các từ mã có kích thước gần với lượng tin riêng của kí tự (sai số ±1)
- Mã Huffman đảm bảo kích thước của bản tin mã hóa nhỏ nhất

symbol	Shannon-Fano Code word	Huffman Code word	Tần suất	Lượng tin riêng -log2p(x)
A	00	0	15	1.38
B	01	100	7	2.48
C	10	101	6	2.7
D	110	110	6	2.7
E	111	111	5	2.96

- Kích thước bản tin

$$L_{Shannon} = 2bit \times (15+7+6) + 3bit \times (6+5) = 89bit \qquad R_{Shannon} = 89bit / 39 = 2.28bit / symbol$$

$$L_{Huffman} = 1bit \times 15 + 3bit \times (7+6+6+5) = 87bit \qquad R_{Huffman} = 87bit / 39 = 2.23bit / symbol$$

H(X)=2.1858

Vi du