CHƯƠNG 4. MÃ HÓA NGUỒN

1.
MỘT SỐ KHÁI NIỆM

Mã hóa nguồn

- Nhiệm vụ của bộ mã hóa nguồn là chuyển đổi nguồn thành một chuỗi các số nhị phân (bit) được gọi là chuỗi thông tin.
- Nếu nguồn là nguồn liên tục, nó sẽ bao gồm quá trình biến đổi A/D.
- Một bộ mã hóa nguồn lý tưởng có thuộc tính sau:
  - Tỷ lệ bit trung bình được yêu cầu để hiển thị đầu ra nguồn cần phải được tối thiểu hóa bằng cách giảm độ dư thừa của nguồn thông tin.
  - Đầu ra nguồn phải có thể tái tạo lại từ chuỗi thông tin mà không có sự sai sót nào.
- Mã hóa nguồn được sử dụng chủ yếu để nén dữ liệu, ví dụ thoại, ảnh, video, text…

Các tham số mã hóa

- Tập tin nguồn: Nguồn thông tin rời rạc gồm một tập hữu hạn các ký tự nguồn là các đầu ra có thể có. Tập các ký tự nguồn này được gọi là tập tin nguồn.
- Ký tự: là các thành phần của một tập tin nguồn.
- Từ mã nhị phân: Đây là sự kết hợp của một số các số nhị phân (bit) được gán cho mỗi ký tự.
- Độ dài từ mã: Số lượng các bit trong một từ mã được gọi là độ dài từ mã.

Các tham số mã hóa

- Độ dài trung bình từ mã: Xét một nguồn rời rạc không nhớ X có entropy hữu hạn $H(X)$ và tập tin $\{x_1, x_2, \dots, x_m\}$ với xác suất xảy ra tương ứng là $P(x_j)$ .Nếu từ mã nhị phân được ấn định cho ký tự x_j có độ dài n_j bit, độ dài trung bình từ mã L cho mỗi ký tự nguồn được định nghĩa là:

$$L = \sum_{j=1}^{m} P(x_j)\, n_j \quad \text{(bit/ký tự)}$$

- Hiệu suất mã: Hiệu suất mã được cho bởi: $\eta = \frac{L_{min}}{L}$
- Ở đó L_{min} là giá trị nhỏ nhất có thể có của L.
- Độ dư thừa mã: Độ dư thừa γ của mã được định nghĩa là: $\gamma = 1 - \eta$

Định lý mã hóa nguồn

- Định lý mã hóa nguồn phát biểu rằng nếu X là một nguồn rời rạc không nhớ với entropy $H(X)$, độ dài từ mã trung bình L trên mỗi ký tự được giới hạn bởi:

$$L \geq H(X)$$

Khi $L_{min} = H(X);\ \eta = \frac{H(X)}{L}$

Ví dụ 4.1: Một nguồn rời rạc không nhớ X gồm hai ký tự x_1 và x_2. Xác suất xảy ra và bộ mã tương ứng được mô tả trong bảng dưới đây. Tìm hiệu suất mã và độ dư thừa mã.

x_j	$P(x_j)$	Từ mã
x_1	0,8	0
x_2	0,2	1

Giải

Độ dài từ mã trung bình trên một ký tự là:

$L = \sum_{j=1}^{N} P(x_j) n_j = 0{,}8.1 + 0{,}2.1 = 1$ (bit)

Entropy là:

$H(X) = -\sum_{j=1}^{2} P(x_j)\, log_2 P(x_j) = -0{,}8.log0{,}8 - 0{,}2log0{,}2 = 0{,}722$ (bit/ký tự)

Hiệu suất mã là: $\eta = \frac{H(X)}{L} = 0{,}722 = 72{,}2\%$

Độ dư thừa mã là: $\gamma = 1 - \eta = 1 - 0{,}722 = 0{,}278 = 27{,}8\%$

x_j	$P(x_j)$	Từ mã
x_1	0,8	0
x_2	0,2	1

Phân loại mã

- Mã độ dài cố định: Nếu độ dài từ mã của một bộ mã là không đổi thì mã được gọi là mã chiều dài cố định. Mã có độ dài cố định ấn định một số bit không đổi cho các ký tự nguồn bất kể xác suất xuất hiện của nó.
  - Ví dụ: Mã ASCII ấn định 7 bit cho mỗi từ mã.
  - Xét một nguồn rời rạc không nhớ gồm m ký tự. Nếu m là lũy thừa của 2, số bit yêu cầu cho mỗi ký tự là $log_2 m$. Nếu m không là lũy thừa của 2, số bit yêu cầu là $[log_2 m+1]$
- Mã độ dài thay đổi: là mã có các độ dài từ mã không cố định.
  - Ví dụ: ấn định số bit từ mã cho các ký tự nguồn tùy thuộc xác suất xuất hiện của ký tự đó.
- Mã phân biệt: là bộ mã mà có thể phân biệt các từ mã với nhau.
  - Ví dụ:

x_j	x_1	x_2	x_3	x_4
Từ mã	00	01	10	11

Phân loại mã

- Mã có khả năng giải mã duy nhất: là bộ mã cho phép tái tạo một cách hoàn hảo các ký tự nguồn từ chuỗi nhị phân đã mã hóa.
- Ví dụ: Mã hóa bản tin **ABADCAB** bằng 2 bộ mã.
- **Mã 1: 00010011100001 (14 bit)**
- **Mã 2: 010010001 (9 bit)**
- Khi giải mã gặp vấn đề gì?

Symbol	Code 1	Code 2
A	00	0
B	01	1
C	10	00
D	11	01

- Mã 1 có thể giải mã ra 1 bản tin duy nhất, còn mã 2 thì không. Mã 2 không có khả năng giải mã duy nhất.
- [0][1][0][0][1][0][0][01] ⟹ ABAABAAD
- [0][1][00][1][00][01] ⟹ ABCBCD

Phân loại mã

- Mã tiền tố: là bộ mã trong đó không có từ mã nào là phần đầu của một từ mã khác.
- Ví dụ mã tiền tố:

Symbol	Code Word
A	0
B	10
C	110
D	1110

- Mã tức thời: Một mã có khả năng giải mã duy nhất được gọi là mã tức thời nếu có thể nhận ra kết thúc của bất cứ từ mã nào mà không cần kiểm tra các ký hiệu từ mã tiếp theo. Bởi vì mã tức thời cũng có thuộc tính là không có từ mã nào là phần đầu của một từ mã khác, nên mã tiền tố cũng được gọi là mã tức thời.
- Mã hóa entropy: Khi bộ mã chiều dài thay đổi được thiết kế sao cho độ dài từ mã trung bình tiếp cận entropy của nguồn rời rạc không nhớ thì mã được gọi là mã entropy. Mã Shannon-Fano và mã Huffman (sẽ thảo luận ở phần sau) là hai ví dụ của loại mã này.
- Mã tối ưu: Một mã được gọi là mã tối ưu nếu nó là mã tức thời và có độ dài từ mã trung bình tối thiểu L đối với một nguồn cho trước với xác suất cụ thể được ấn định cho các ký tự nguồn.

Ví dụ 4.2

Xét hai bộ mã nhị phân có 4 từ mã. So sánh hiệu quả của hai bộ mã.

x_j	$P(x_j)$	Code 1	Code 2
x_1	0.5	00	0
x_2	0.25	01	10
x_3	0.125	10	110
x_4	0.125	11	111

Giải

- Bộ mã 1 là mã có độ dài cố định là 2.
- Độ dài từ mã trung bình của từ mã là:
- $L = \sum_{j=1}^{4} P(x_j) n_j$
- $= 0{,}5.2 + 0{,}25.2 + 0{,}125.2 + 0{,}125.2$
- $= 2\ (bits)$
- Entropy là:
- $H(X) = \sum_{j=1}^{4} P(x_j) log P(x_j)$= 1,75 bit/ký tự
- Hiệu suất mã là:
- $\eta = \frac{H(X)}{L} = \frac{1{,}75}{2}.100\% = 87{,}5\%$

- Bộ mã 2 là mã có độ dài thay đổi.
- Độ dài từ mã trung bình của từ mã là:
- $L = \sum_{j=1}^{4} P(x_j) n_j$
- $= 0{,}5.1 + 0{,}25.2 + 0{,}125.3 + 0{,}125.3$
- $= 1{,}75\ (bits)$
- Entropy là:
- $H(X) = \sum_{j=1}^{4} P(x_j) log P(x_j)$= 1,75 bit/ký tự
- Hiệu suất mã là:
- $\eta = \frac{H(X)}{L} = \frac{1{,}75}{1{,}75}.100\% = 100\ \%$

Bất đẳng thức Kraft

Gọi X là nguồn rời rạc không nhớ có các ký tự $\{x_j\}$ $(j = 1,2, \dots, m)$. Nếu độ dài từ mã nhị phân tương ứng với x_j là n_j thì điều kiện cần và đủ để tồn tại mã nhị phân tức thời là:

$$K = \sum_{j=1}^{m} 2^{-n_j} \leq 1$$

Đây là bất đẳng thức Kraft. Bất đẳng thức cho biết sự tồn tại của mã có khả năng giải mã tức thời với độ dài từ mã thỏa mãn bất đẳng thức.

Ví dụ 4.3

Xét một nguồn rời rạc không nhớ với bốn ký tự nguồn được mã hóa bằng các bộ mã nhị phân khác nhau. Chứng minh rằng:

(a) Các bộ mã 1, 3, 4 thỏa mãn bất đẳng thức Kraft.

(b) Các bộ mã 1, 4 là mã có khả năng giải mã duy nhất còn các bộ mã 2,3 không phải là mã có khả năng giải mã duy nhất.

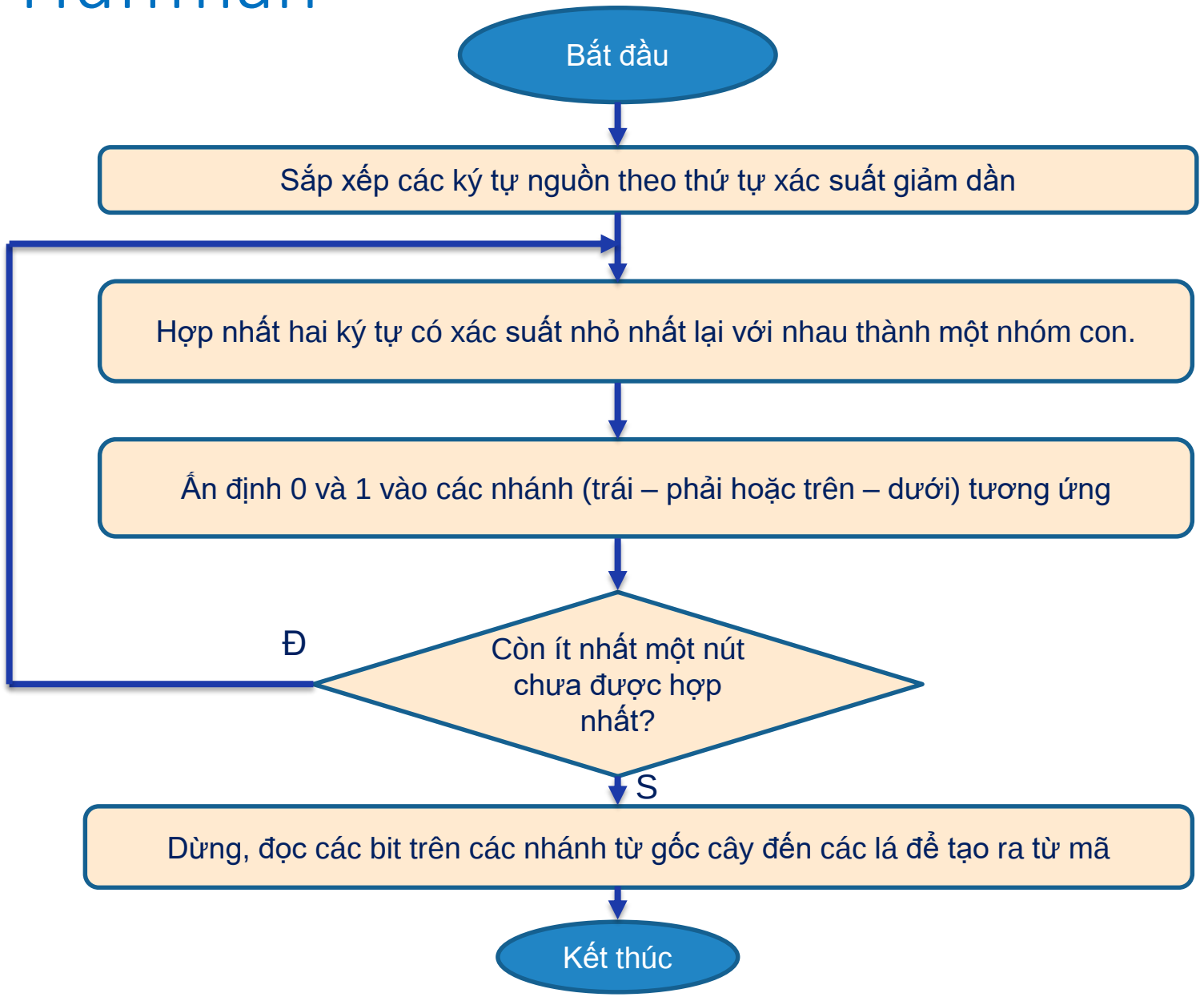
x_j	Code 1	Code 2	Code 3	Code 4
x_1	00	0	0	0
x_2	01	10	11	100
x_3	10	11	100	110
x_4	11	110	110	111

2.
MÃ HUFFMAN

Mã hóa Huffman

- Mã hóa Huffman luôn tạo ra các mã tiền tố có độ dài từ mã trung bình nhỏ nhất có thể. Đây là mã tối ưu có hiệu suất cao nhất hay mã có độ thừa nhỏ nhất. Vì vậy nó còn được gọi là mã có độ dư thừa nhỏ nhất hay mã tối ưu.
- Mã Huffman được sử dụng trong JPEG, MPEG-1/2/4, H.261, H.262, H.263, H.264…

Thuật toán Huffman



Ví dụ 4.4

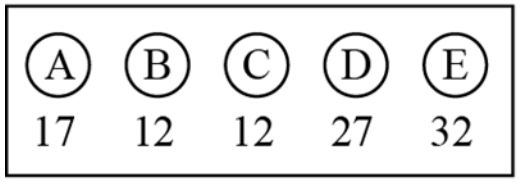
Cho tập 5 ký tự A, B, C, D, E với tần suất xuất hiện tương ứng trong bảng dưới đây.

A. Thực hiện mã hoá Huffman cho tập ký tự này.
B. Tính hiệu suất mã

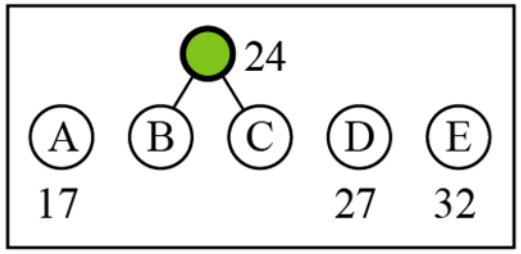
Ký tự	**A**	**B**	**C**	**D**	**E**
Tần suất xuất hiện	17	12	12	27	32
	0,17	0,12	0,12	0,27	0,32

Giải

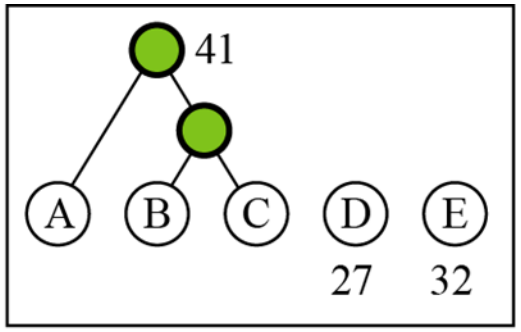
Ký tự	A	B	C	D	E
Tần suất xuất hiện	17	12	12	27	32



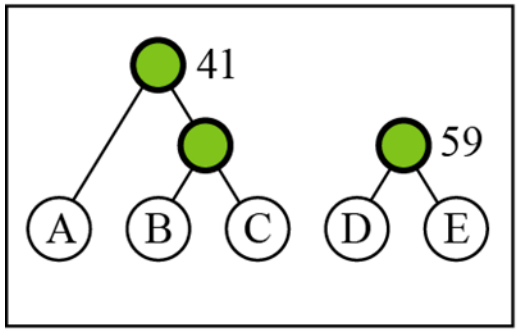
a.



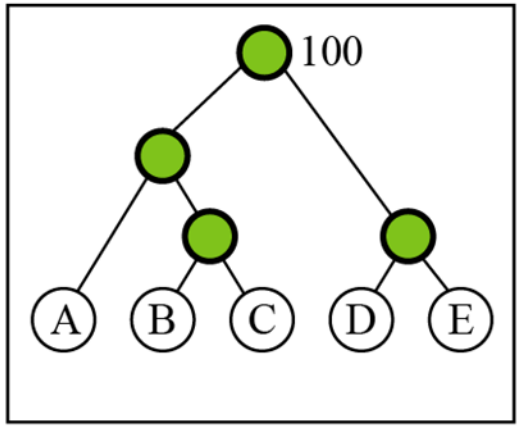
b.



c.

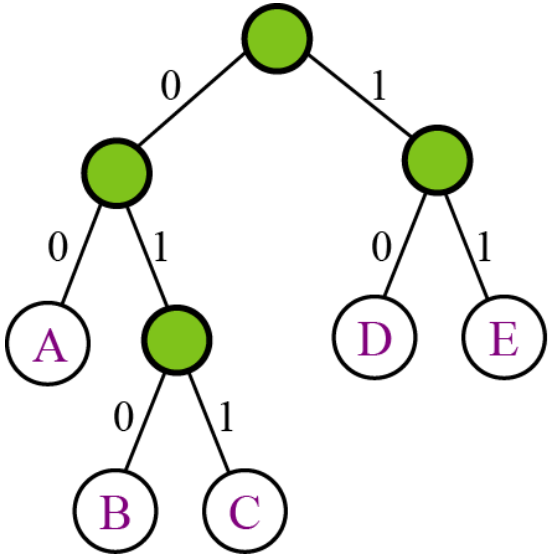


d.



e.

Giải

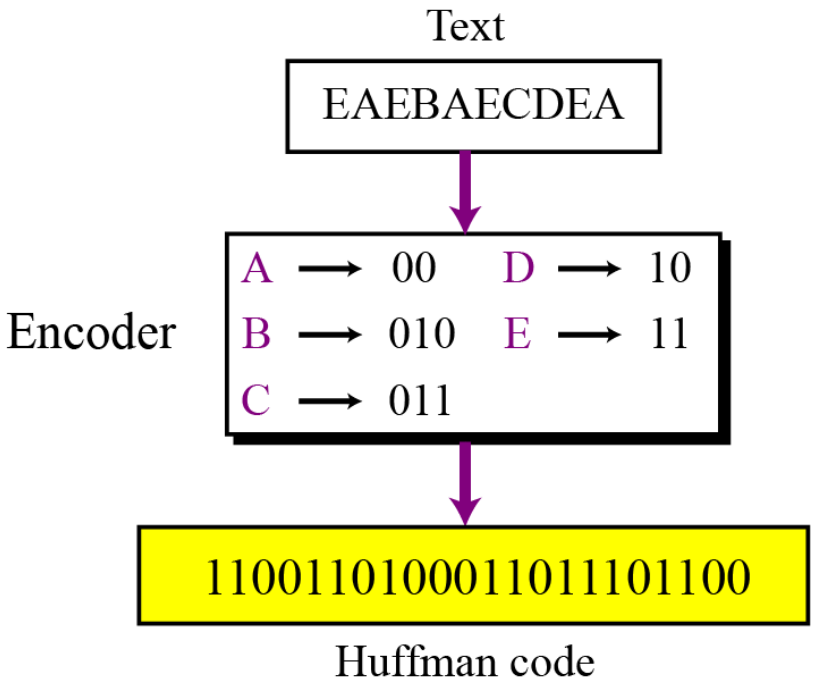


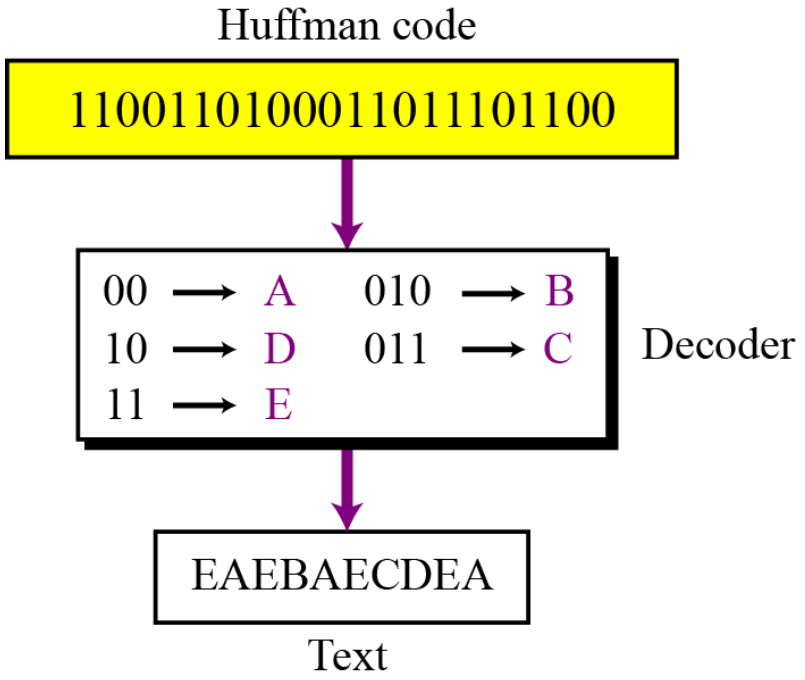
A: 00				D: 10	
B: 010				E: 11	
C: 011					

Code

Ký tự	**A**	**B**	**C**	**D**	**E**
Tần suất xuất hiện	17	12	12	27	32

Ví dụ quá trình mã hóa và giải mã





Bài tập

- Cho tập tin gồm 5 ký tự A, B, C, D, E với các xác suất lần lượt là: 0,1; 0,2; 0,1; 0,15 và 0,45.
- (a) Hãy xây dựng hai cây mã hóa Huffman cho nguồn trên.
- (b) Đánh giá hiệu quả của hai bộ mã

Bài tập

1) Một văn bản được viết từ các ký tự từ x1 ÷ x14, biết tần suất xuất hiện của các ký tự trong văn bản lần lượt là: 1200; 2400; 9600; 2400; 9600; 2400; 1200; 9600; 9600; 38400; 9600; 9600; 9600; 38400 (lần).

a. Hãy thực hiện mã hóa Huffman cho văn bản.

b. Đánh giá hiệu quả của phép mã hóa xây dựng trong câu a.

c. Kiểm tra bất đẳng thức kẹp về độ dài trung bình từ mã. Có nhận xét gì?

d. Hãy tính tỷ số nén thu được khi sử dụng bộ mã xây dựng ở phần a so với khi sử dụng mã ASCII với độ rộng 1 Byte.

2) Thực hiện mã hóa Huffman và tính độ dài trung bình từ mã cho nguồn rời rạc sau:

A = (a1,a2,a3,a4,a5,a6) có xác suất lần lượt là: (0,12; 0,08; 0,3; 0,15;0,3; 0,05)

3) Yêu cầu tương tự bài 2) với A = (a1, a2,a3, a4, a5, a6, a7) có xác suất lần lượt

là: (1/32; 1/16; 1/8;1/32;1/4;1/4;1/4).

Giải mã cho chuỗi dữ liệu nhận được 000111001011…