Лабораторне заняття 5.

Тема: Дослідження алгоритмів ефективного кодування.

Мета: дослідити процес кодування даних за допомогою алгоритмів ефективного кодування.

***Завдання**: розібрати приклади 1 та 2, виконати аналогічні розрахунки і побудови*

Приклад 1.

Джерело генерує знак x_1 с ймовірністю 0,8 і x_2 з ймовірністю 0,2. Побудувати ефективні коди Шеннона-Фано і Хафмана для послідовностей з трьох знаків $x_i\, x_j\, x_k$. Яке середнє число символів на знак? Порівняти з ентропією джерела.

Розв'язок.

Ентропія джерела за формулою Шеннона H(x)=0,722. Для найбільш наочного порівняння ентропії джерела із середнім значенням довжини кодового позначення розглянемо різні комбінації знаків. Застосуємо метод Шеннона-Фано.

знак	вірогідність	код
x_1	0,8	1
x_2	0,2	0

Середня кількість символів в коді (або довжина)
$n_{\text{ср}}$ =1*0,8+1*0,2=1.

Середнє число символів на знак $\frac{n_{\text{ср}}}{1}$ = 1, що на 28% більше ентропії.

Побудуємо код Шеннона-Фано для різних двох знакових комбінацій.

Комбінація знаків	Вірогідність	Символи коду			код
		1-й	2-й	3-й	
x_1x_1	0,64	1			1
x_1x_2	0,16	0	1		01
x_2x_1	0,16		0	1	001
x_2x_2	0,04			0	000

Середня кількість символів в коді (або довжина)
$n_{\text{ср}} = 1 * 0{,}64 + 2 * 0{,}16 + 3 * 0{,}16 + 3 * 0{,}04 = 1{,}56$

Середнє число символів на знак $\frac{n_{\text{ср}}}{2}$ = 0,78 (знаменник означає з скількох знаків полягає одна комбінація), що на 8% більше ентропії.

Побудуємо код Шеннона-Фано для різних трьох знакових комбінацій.

Комбінація знаків	Вірогідність	Символи коду					код
		1-й	2-й	3-й	4-й	5-й	
$x_1x_1x_1$	0,512	1					1
$x_1x_1x_2$	0,128	0	1	1			011
$x_1x_2x_1$	0,128			0			010
$x_2x_1x_1$	0,128		0	1			001
$x_1x_2x_2$	0,032			0	1	1	00011
$x_2x_1x_2$	0,032					0	00010
$x_2x_2x_1$	0,032				0	1	00001
$x_2x_2x_2$	0,008					0	00000

Середня кількість символів в коді (або довжина)

$$n_{\text{ср}} = 1 * 0{,}512 + 3 * 0{,}384 + 3 * 0{,}16 + 5 * 0{,}104 = 2{,}184,$$

Середнє число символів на знак $\frac{n_{\text{ср}}}{3} = 0{,}728$ (знаменник означає зі скількох знаків складається одна комбінація), що на 8% більше ентропії.

Побудуємо код Хафмана для всіляких трьох знакових комбінацій, рис. 1:

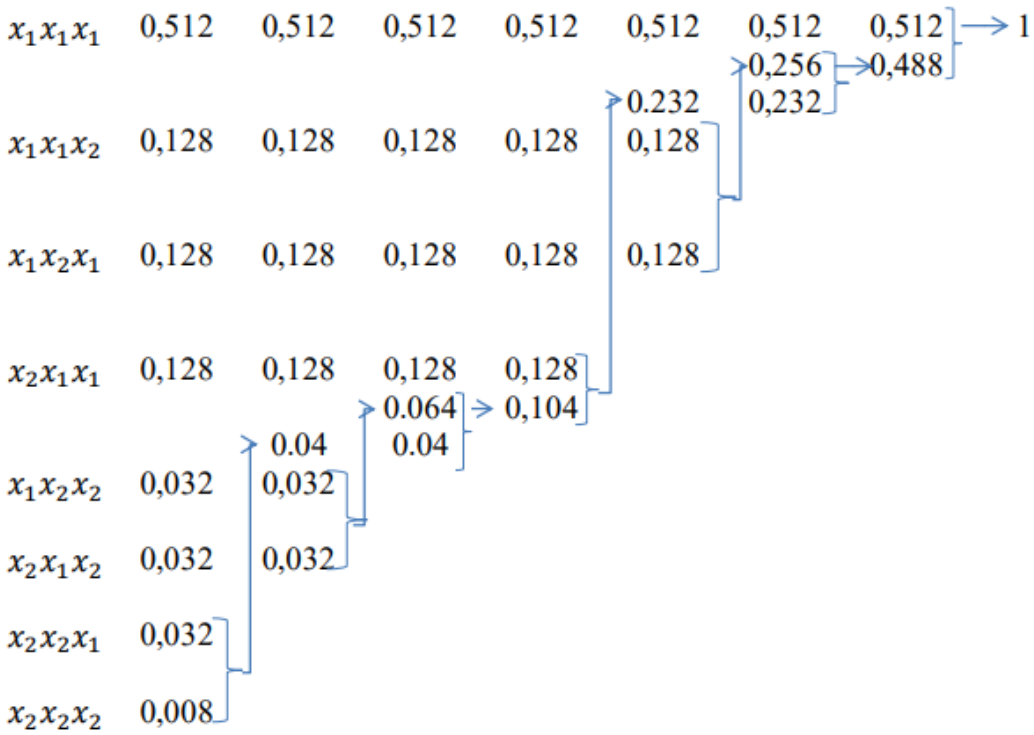


Рис.1. Код Хаффмена

Будуємо кодове дерево, рис. 2:

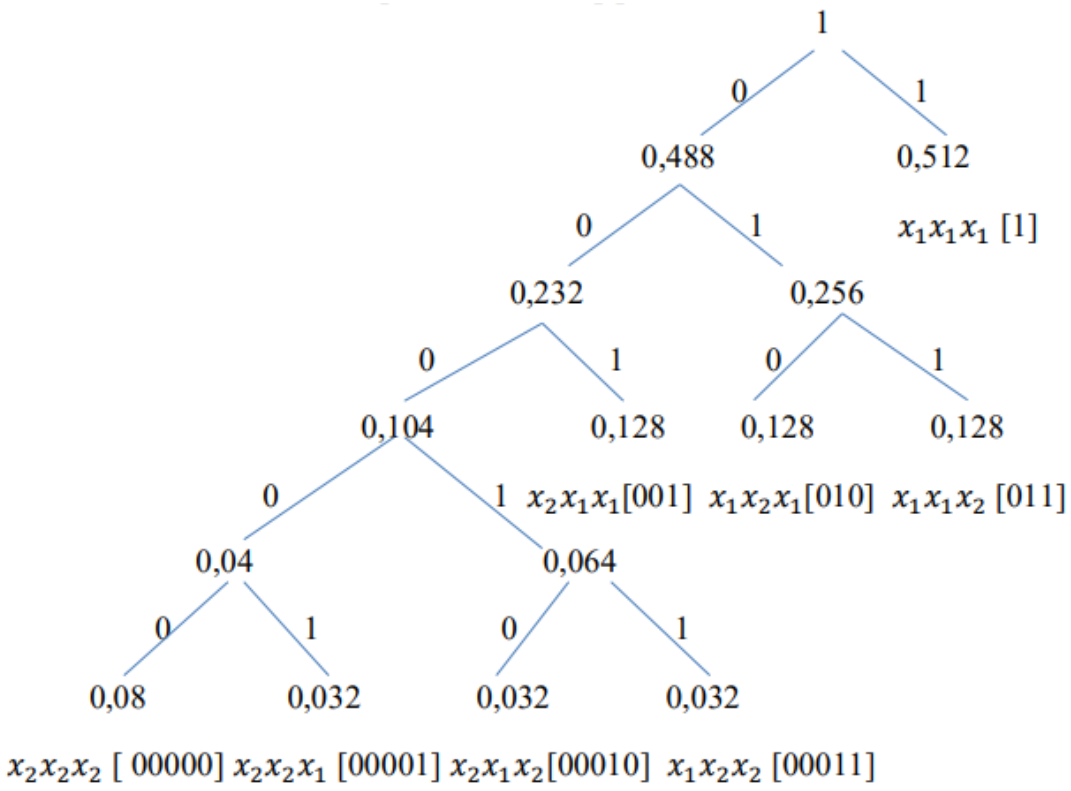


Рис. 2. Кодове дерево Хаффмена

Складаємо код комбінацій:

$x_1x_1x_1$	1
$x_1x_1x_2$	011
$x_1x_2x_1$	010
$x_2x_1x_1$	001
$x_1x_2x_2$	00011
$x_2x_1x_2$	00010
$x_2x_2x_1$	00001
$x_2x_2x_2$	00000

Коди Шеннона-Фано і Хаффмена збіглися.
Середня кількість символів в коді (або довжина)
$n_{\text{ср}} = 1 * 0{,}512 + 3 * 0{,}384 + 3 * 0{,}16 + 5 * 0{,}104 = 2{,}184,$

Середнє число символів на знак $\frac{n_{\text{ср}}}{3}$ = 0,728 (знаменник означає з скількох знаків полягає одна комбінація), що на 8% більше ентропії.

Приклад 2.

Повідомлення X з символами x_1, x_2, x_3, x_4, x_5 передається по дискретному бінарного каналу з можливостями $p(x_1)$ = 0,1 , $p(x_2)$ = 0,1 , $p(x_3)$ = 0,4 , $p(x_4)$ = 0,3 , $p(x_5)$ = 0,1. Смуга пропускання каналу забезпечує можливість передачі двійкових символів тривалістю $\tau = 10^{-4}$сек. Потрібно вибрати найкращий спосіб кодування.

Рішення.

Пропускна здатність двійкового дискретного каналу $\text{C} = \frac{1}{\tau}$ = 1000 біт/сек.

1) Повідомлення закодовано рівномірним двійковим кодом.

X	P(X)	код
x_3	0,4	000
x_4	0,3	001
x_1	0,1	010
x_2	0,1	011
x_5	0,1	111

Середня кількість символів в коді (або довжина)

$n_{\text{ср}} = 3 * 0{,}4 + 3 * 0{,}3 + 3 * 0{,}1 + 3 * 0{,}1 + 3 * 0{,}1 = 3,$

$\tau_{\text{ср}} = n_{\text{ср}} * \tau = 3 * 0{,}0001 = 0{,}0003,$

$H(X) = -\sum_{i=1}^{n} p(x_i) log_2(x_i) = -(0{,}4 log_2 0{,}4 + 0{,}3 log_2 0{,}3 + 3 * 0{,}1 log_2 0{,}1) = 2{,}046$ біт

$V = \frac{H(X)}{\tau_{\text{ср}}} = \frac{2{,}046}{0{,}0003} = 6820$

2) Повідомлення закодовано кодом Шеннона - Фано.

X	P(X)	код
x_3	0,4	0
x_4	0,3	10
x_1	0,1	110
x_2	0,1	1110
x_5	0,1	1111

Середня кількість символів в коді (або довжина)

$n_{\text{ср}} = 1 * 0{,}4 + 2 * 0{,}3 + 3 * 0{,}1 + 4 * 0{,}1 + 4 * 0{,}1 = 2{,}1,$

$\tau_{\text{ср}} = n_{\text{ср}} * \tau = 2{,}1 * 0{,}0001 = 0{,}00021,$

$$V = \frac{H(X)}{\tau_{\text{ср}}} = \frac{2{,}046}{0{,}00021} = 9742{,}86$$

3) Повідомлення закодовано кодом Хаффмена.

X	P(X)	код
x_3	0,4	0
x_4	0,3	11
x_1	0,1	100
x_2	0,1	1011
x_5	0,1	1010

Середня кількість символів в коді (або довжина)

$n_{\text{ср}} = 1 * 0{,}4 + 2 * 0{,}3 + 3 * 0{,}1 + 4 * 0{,}1 + 4 * 0{,}1 = 2{,}1,$

$\tau_{\text{ср}} = n_{\text{ср}} * \tau = 2{,}1 * 0{,}0001 = 0{,}00021,$

$$V = \frac{H(X)}{\tau_{\text{ср}}} = \frac{2{,}046}{0{,}00021} = 9742{,}86$$

Найповільнішим кодом виявився рівномірний код V =6820 сим/сек, що в 1,43 рази менше, ніж при кодуванні методом Шеннона - Фано і Хаффмена.

Завдання 1.

Побудувати коди: рівномірний, Шеннона-Фано та Хаффмена (кодове дерево) для наступних даних:

A	B	C	D	E	F	G	H
0,25	0,22	0,13	0,11	0,1	0,09	0,07	0,03

Оцінити середню кількість символів і порівняти. Зробити висновок.

Завдання 2.

Побудувати коди: рівномірний, Шеннона-Фано та Хаффмена (кодове дерево) для наступних даних: Ваше *Ім'я _ прізвище*. Сума вибраних Вами імовірностей генерації символів повинна складати одиницю. Оцінити ентропію джерела повідомлення, середню кількість символів і порівняти. Зробити висновок.

Відносні частоти вживання літер української абетки.

Літера	Частота	Літера	Частота	Літера	Частота	Літера	Частота
„–”	0,138	**Р**	0,043	**П**	0,025	**Ї**	0,010
О	0,086	**Е**	0,042	**З**	0,020	**Ц**	0,010
Н	0,068	**С**	0,037	**Я**	0,019	**Й**	0,009
А	0,064	**К**	0,033	**Ь**	0,016	**Ю**	0,008
И	0,055	**М**	0,029	**Б**	0,013	**Ж**	0,007
В	0,046	**У**	0,027	**Г**	0,013	**Ш**	0,005
Т	0,045	**Д**	0,027	**Ч**	0,011	**Є**	0,005
І	0,044	**Л**	0,027	**Х**	0,011	**Щ**	0,004
						Ф	0,003

Частота повторюваності букви Ґ в текстах українською мовою дорівнює $6*10^{-5}$