Шенон - Фаноов префиксни код

- Шенонова теорема кодовања (1. дио): $E[L] \geq \frac{H(S)}{\log_2 D}$
- Шенон - Фаноов код, иако није оптималан, у општем случају се довољно добро приближава доњој граници очекиване вриједности дужине кодних ријечи
- идеја се своди на избор дужине l_i кодне ријечи s_i тако да важи:

$$l_i = \left\lceil -\frac{\log_2 p_i}{\log_2 D} \right\rceil$$

- Шенонова теорема кодовања (2. дио): $E[L] < \frac{H(S)}{\log_2 D} + 1$
- поступак за случај бинарног кода:

  1° симболи са листе се поређају према опадајућем редослиједу $p_1 \geq p_2 \geq \dots \geq p_q$

  2° скуп симбола се дијели на два подједнако или приближно подједнако вјероватна дијела

  3° кодне ријечи за симболе једног подскупа почињу са 0, а другог подскупа са 1

  4° сваки од ових подскупова се даље дијели по истом принципу, додајући на одговарајуће мјесто у кодној ријечи 0 или 1

(1.) Дат је дискретни извор без меморије S са симболима s_1, s_2, s_3 и s_4 и одговарајућим вјероватноћама појављивања p_1, p_2, p_3 и p_4.
Шеноновим поступком одредити кодне ријечи и утврдити да ли је добијени код компактан.

S	p_i	1. подјела	2. подјела	3. подјела	
s_1	0,5	0	0	0	s_1 (0)
s_2	0,25	1	10	10	s_2 (10)
s_3	0,125	1	11	110	s_3 (110)
s_4	0,125	1	11	111	s_4 (111)

$$H(S) = -\sum_{i=1}^{4} p_i \log_2 p_i = -(0{,}5\log_2 0{,}5 + 0{,}25\log_2 0{,}25 + 2\cdot 0{,}125\log_2 0{,}125) = 1{,}75 \frac{sh}{simb}$$

$$\bar{L} = 1\cdot 0{,}5 + 2\cdot 0{,}25 + 2\cdot 3\cdot 0{,}125 = 0{,}5 + 0{,}5 + 0{,}75 = 1{,}75 \frac{b}{simb} \Rightarrow$$ добијени код је компактан

2.

S	p_i	I	II	III		
s_1	0,6	0	0	0	0	s_1 (0)
s_2	0,2	1	10	10	10	s_2 (10)
s_3	0,1	1	11	110	110	s_3 (110)
s_4	0,07	1	11	111	1110	s_4 (1110)
s_5	0,03	1	11	111	1111	s_5 (1111)

$$H(S) = -\sum_{i=1}^{5} p_i \log_2 p_i = 1{,}66 \frac{sb}{simb}$$

$$\overline{L} = 1\cdot 0{,}6 + 2\cdot 0{,}2 + 3\cdot 0{,}1 + 4\cdot 0{,}07 + 4\cdot 0{,}03 = 1{,}7 \frac{b}{simb}$$

3.

S	p_i	I	II	III	
s_1	0,3	0	00	00	s_1 (00)
s_2	0,2	0	01	01	s_2 (01)
s_3	0,2	1	10	10	s_3 (10)
s_4	0,2	1	11	110	s_4 (110)
s_5	0,1	1	11	111	s_5 (111)

$$H(S) = -\sum_{i=1}^{5} p_i \log_2 p_i = 2{,}25 \frac{sb}{simb}$$

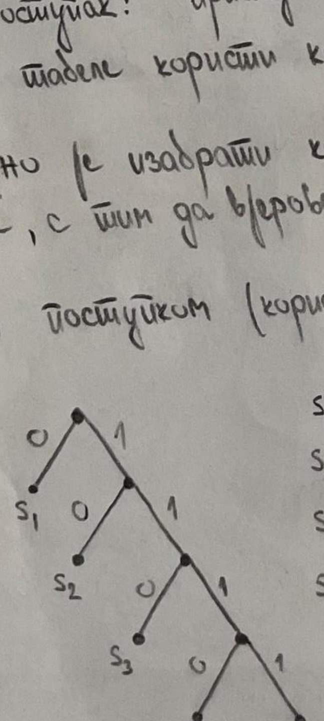
$$\overline{L} = 2\cdot(0{,}3+0{,}2+0{,}2) + 3\cdot(0{,}2+0{,}1) = 2{,}3 \frac{b}{simb}$$

- Фаноов поступак: приступ је сличан као код Шенонови поступка, само се уместо табеле користи кодно стабло.

Потребно је изабрати кодно стабло и завршним чворовима придружити симболе, с тим да вероватноће грана у сваком чвору буду приближно подједнаке.

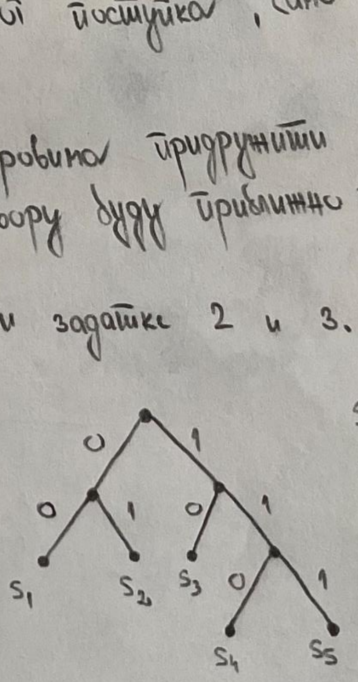
4.) Фановим поступком (користећи кодно стабло) решити задатке 2 и 3.

s_i	p_i
s_1	0,6
s_2	0,2
s_3	0,1
s_4	0,07
s_5	0,03



$s_1 \to 0$
$s_2 \to 10$
$s_3 \to 110$
$s_4 \to 1110$
$s_5 \to 1111$

s_i	p_i
s_1	0,3
s_2	0,2
s_3	0,2
s_4	0,2
s_5	0,1



$s_1 \to 00$
$s_2 \to 01$
$s_3 \to 10$
$s_4 \to 110$
$s_5 \to 111$

Код Шенон - Фаноовог поступка може да се генерише више различитих кодова (кодних стабала) зависно од расподјеле вјероватноћа.
У овом случају требало би испитати сва могућа кодна стабла (има их коначан број) а потом изабрати стабло које даје најмању средњу дужину кодних ријечи.
Дакле једнострука примјена овог поступка не гарантује добијање компактног кода.

* Нека је дат извор са $q=5$ симбола. Могућа су три различита бинарна кода.

s_i	p_i	a)	b)	c)
s_1	0,6	0	0	00
s_2	0,2	10	100	01
s_3	0,1	110	101	10
s_4	0,07	1110	110	110
s_5	0,03	1111	111	111

$H(S) = 1{,}66 \frac{sh}{simb}$

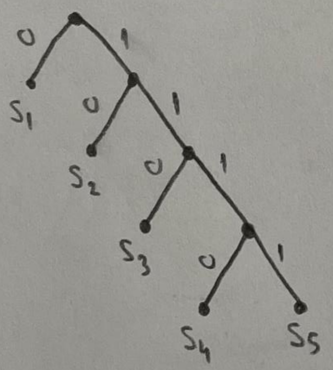
$\bar{L}_a = 1{,}7 \frac{b}{simb}$

$\bar{L}_b = 1{,}8 \frac{b}{simb}$

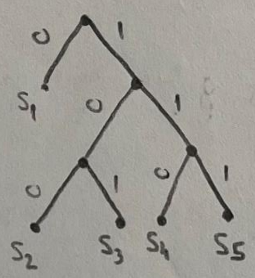
$\bar{L}_c = 2{,}1 \frac{b}{simb}$

$\Rightarrow \bar{L}_a$ обезбјеђује компактни код

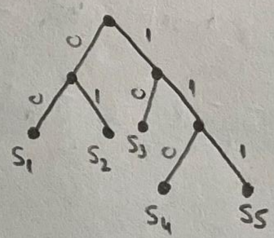
a)



b)



c)



Хафманов поступак

- Хафманов алгоритам кодовања извора информација без меморије, помоћу D-арног префиксног кода, даје оптималан код минималне могуће очекиване вриједности дужине кодних ријечи

1° вјероватноће се поредају у опадајућем редослиједу

2° одреди се највећи цијели број g_0 $(2 \leq g_0 \leq m)$ који задовољава услов да је $\frac{q - g_0}{m-1}$ цио број (m представља арност кода којим се врши кодовање)

3° формира се подскуп од g_0 симбола најмање вјероватноће

4° формирамо нову листу S_1

5° групишемо m најмање вјероватних симбола

6° формирамо нову листу S_2

Поступак се понавља док не дођемо до m симбола.

1. Дата је изворна листа U са $q = 6$ симбола чије су вјероватноће појављивања дате у табели. Хафмановим поступком конструисати код при чему се кодовање врши бинарним кодом.

U	u_1	u_2	u_3	u_4	u_5	u_6
p_i	0,05	0,1	0,15	0,27	0,2	0,23

U: u_i	$p(u_i)$		U_1		U_2		U_3		U_4	
u_4	0,27	01	0,27	01	0,3	00	0,43	1	0,57	0
u_6	0,23	10	0,23	10	0,27	01	0,3	00	0,43	1
u_5	0,2	11	0,2	11	0,23	10	0,27	01		
u_3	0,15	000	0,15	000	0,2	11				
u_2	0,1	0010	0,15	001						
u_1	0,05	0011								

$u_1 \to 0011$

$u_2 \to 0010$

$u_3 \to 000$

$u_4 \to 01$

$u_5 \to 11$

$u_6 \to 10$

$$H(U) = -\sum_{i=1}^{6} p_i \log_2 p_i = 2{,}42 \ \frac{sh}{simb}$$

$$E(U) = 2\cdot(0{,}27 + 0{,}2 + 0{,}23) + 3\cdot 0{,}15 + 4\cdot(0{,}05 + 0{,}1) =$$

$$= 2{,}45 \ \frac{b}{simb}$$

2. Дата је секвенца слова и њихове вероватноће појављивања.

s_i	s_0	s_1	s_2	s_3	s_4
$p(s_i)$	0,55	0,15	0,15	0,10	0,05

Конструисати 2 различита бинарна Хафманова кода за овај скуп. При конструкцији кода сместити на врх у првом случају комбиноване симболе листе, а у другом на дно листе. У оба случаја израчунати просечну дужину кодних речи и варијансу просечне дужине кодних речи над датим ансамблом слова.

I:

S			S_1		S_2		S_3	
s_i	$p(s_i)$							
s_0	0,55	0	0,55	0	0,55	0	0,55	0
s_1	0,15	100	0,15	11	0,3	10	0,45	1
s_2	0,15	101	0,15	100	0,15	11		
s_3	0,1	110	0,15	101				
s_4	0,05	111						

$$\bar{L} = \sum_{i=0}^{4} p(s_i)\cdot l_i = 0{,}55\cdot 1 + 3\cdot(0{,}15+0{,}15+0{,}1+0{,}05) = 0{,}55 + 1{,}35 = \boxed{1{,}9\,\frac{b}{simb}}$$

$$\sigma^2 = \sum_{i=0}^{4} p(s_i)\cdot(l_i-\bar{L})^2 = 0{,}55\cdot(1-1{,}9)^2 + (0{,}15+0{,}15+0{,}1+0{,}05)(3-1{,}9)^2 =$$
$$= 0{,}55\cdot(-0{,}9)^2 + 0{,}45\cdot(1{,}1)^2 \boxed{= 0{,}99}$$

II:

			S_1		S_2		S_3	
s_i	$p(s_i)$							
s_0	0,55	0	0,55	0	0,55	0	0,55	0
s_1	0,15	11	0,15	11	0,3	10	0,45	1
s_2	0,15	100	0,15	100	0,15	11		
s_3	0,1	1010	0,15	101				
s_4	0,05	1011						

$$\bar{L} = \sum_{i=0}^{4} p(s_i)\cdot l_i = 0{,}55\cdot 1 + 2\cdot 0{,}15 + 3\cdot 0{,}15 + 4(0{,}1+0{,}05) = 0{,}55+0{,}3+0{,}45+0{,}6 = \boxed{= 1{,}9\,\frac{b}{simb}}$$

$$\sigma^2 = \sum_{i=0}^{4} p(s_i)\cdot(l_i-\bar{L})^2 = 0{,}55\cdot(1-1{,}9)^2 + 0{,}15(2-1{,}9)^2 + 0{,}15(3-1{,}9)^2 + (0{,}1+0{,}05)(4-1{,}9)^2 = \boxed{= 1{,}29}$$

$\Rightarrow$ код који има мању варијансу је бољи

3. Дата је изворна листа S са $q=8$ симбола чије су вероватноће појаве $p(s_i)$ дате у табели. Конструисати компактни код Хафмановим поступком при чему се кодовање врши тернарним кодом са кодном листом $\{0,1,2\}$.

$q=8$

$m=3$

$\frac{q-q_0}{m-1}$ треба да је цио број, при чему је $2 \leq q_0 \leq m$

$\frac{8-q_0}{3-1} = \frac{8-q_0}{2} \in \mathbb{Z} \Rightarrow q_0 = 2$

s_i	$p(s_i)$		S_1		S_2		S_3	
s_1	0,28	1	0,28	1	0,28	1	0,46	0
s_2	0,18	00	0,18	00	0,26	2	0,28	1
s_3	0,15	01	0,15	01	0,18	00	0,26	2
s_4	0,13	02	0,13	02	0,15	01		
s_5	0,1	20	0,1	20	0,13	02		
s_6	0,07	22	0,09	21				
s_7	0,05	210	0,07	22				
s_8	0,04	211						

$$E(L) = \sum_{i=1}^{8} p(s_i)\cdot l_i = 0{,}28\cdot 1 + 2\cdot(0{,}18+0{,}15+0{,}13+0{,}1+0{,}07) + 3\cdot(0{,}05+0{,}04) =$$

$$= 0{,}28 + 1{,}26 + 0{,}27 = \boxed{1{,}81 \frac{b}{simb}}$$

$$H(L) = -\sum_{i=1}^{8} p_i \log_2 p_i = \ldots = 2{,}76 \frac{sh}{simb}$$

$$E(L) \geq \frac{H(L)}{\log_2 D} = \frac{H(L)}{\log_2 3} = \frac{2{,}76}{1{,}58} = 1{,}75$$

$$1{,}75 \leq 1{,}81 < 2{,}75$$