

Домашняя работа 1

Данила Печенев

23 марта 2022 г.

Задача 1.

- $\{1, 10, 0, 01\}$ - не префиксный, так как код 1 является началом кода 10;
- $\{00, 010, 011, 01\}$ - не префиксный, так как код 01 является началом кода 011;
- $\{10, 010, 011, 11\}$ - префиксный;
- $\{1, 00, 010, 011\}$ - префиксный.

Задача 2.

- $\{00, 010, 011, 01\}$ - разделимый, так как в нем выполнено условие постфиксности (можем декодировать сообщение с конца);
- $\{1, 10, 0, 01\}$ - не разделимый. Сообщение 10 можно декодировать как ас и как b;
- $\{10, 010, 011, 11\}$ - разделимый, так как в нем выполнено условие префиксности;
- $\{01, 010, 110, 11\}$ - разделимый, так как в нем выполнено условие постфиксности (можем декодировать сообщение с конца).

Задача 3.

Пусть даны символы a, b, c, d: $P(a) = 0.5$, $P(b) = 0.3$, $P(c) = 0.1$, $P(d) = 0.1$.

Шаг 1: a - 1, {b, c, d} - 0;

Шаг 2: a - 1, b - 01, {c, d} - 00;

Шаг 2: a - 1, b - 01, c - 001, d - 000.

Средняя длина кода (матожидание) $= 0.5 * 1 + 0.3 * 2 + 0.1 * 3 + 0.1 * 3 = 1.7$

Задача 4.

1) Количество каждого символа из алфавита (Λ) в рассматриваемой строке:

a - 25, b - 34, c - 25, d - 21, $\langle \text{space} \rangle$ - 3, $\langle \text{EOF} \rangle$ - 1.

Всего символов в строке - 108 $\Rightarrow P(a) = \frac{25}{108}$, $P(b) = \frac{34}{108}$, $P(c) = \frac{25}{108}$, $P(d) = \frac{21}{108}$, $P(\langle \text{space} \rangle) = \frac{2}{108}$, $P(\langle \text{EOF} \rangle) = \frac{1}{108}$.

$H = - \sum_{s \in \Lambda} P(s) \cdot \log_2 P(s) = 2.13$ бит.

2) Код Шеннона.

Исходный набор:

Символ	a	b	c	d	<space>	<EOF>
P	$\frac{25}{108}$	$\frac{34}{108}$	$\frac{25}{108}$	$\frac{21}{108}$	$\frac{2}{108}$	$\frac{1}{108}$

Сортируем элементы алфавита по невозрастанию P:

Символ	b	a	c	d	<space>	<EOF>
P	$\frac{34}{108}$	$\frac{25}{108}$	$\frac{25}{108}$	$\frac{21}{108}$	$\frac{2}{108}$	$\frac{1}{108}$

В получившейся таблице каждому символу сопоставляем сумму вероятностей символов до него L:

Символ	b	a	c	d	<space>	<EOF>
L	0	$\frac{34}{108}$	$\frac{59}{108}$	$\frac{84}{108}$	$\frac{105}{108}$	$\frac{107}{108}$

Переведем L в двоичную систему счисления:

Символ	b	a	c	d	<space>	<EOF>
L	0.0000	0.01010	0.10001	0.11000	0.11111001	0.111111011

Посчитаем $C = \lceil -\log_2 P \rceil$ и запишем коды

Символ	b	a	c	d	<space>	<EOF>
C	2	3	3	3	6	7
Код	00	010	100	110	111101	1111110

Код Шеннона-Фано.

Шаг 1: {b, a} - 1, {c, d, <space>, <EOF>} - 0;

Шаг 2: b - 11, a - 10, c - 01, {d, <space>, <EOF>} - 00;

Шаг 3: b - 11, a - 10, c - 01, d - 001, {<space>, <EOF>} - 000;

Шаг 4: b - 11, a - 10, c - 01, d - 001, <space> - 0001, <EOF> - 0000.

Код Хаффмана.

Шаг 1: $P(a) = \frac{25}{108}$, $P(b) = \frac{34}{108}$, $P(c) = \frac{25}{108}$, $P(d) = \frac{21}{108}$, $P(<space>) = \frac{2}{108}$, $P(<EOF>) = \frac{1}{108}$.

Шаг 2: $P(a) = \frac{25}{108}$, $P(b) = \frac{34}{108}$, $P(c) = \frac{25}{108}$, $P(d) = \frac{21}{108}$, $P(<space><EOF>) = \frac{3}{108}$.

Шаг 3: $P(a) = \frac{25}{108}$, $P(b) = \frac{34}{108}$, $P(c) = \frac{25}{108}$, $P(d<space><EOF>) = \frac{24}{108}$.

Шаг 4: $P(a) = \frac{25}{108}$, $P(b) = \frac{34}{108}$, $P(cd<space><EOF>) = \frac{49}{108}$.

Шаг 5: $P(ab) = \frac{59}{108}$, $P(cd<space><EOF>) = \frac{49}{108}$.

Шаг 6: ab - 1, cd<space><EOF> - 0.

Шаг 7: a - 11, b - 10, c - 01, d<space><EOF> - 00.

Шаг 8: a - 11, b - 10, c - 01, d - 001, <space><EOF> - 000.

Шаг 8: a - 11, b - 10, c - 01, d - 001, <space> - 0001, <EOF> - 0000.

3) Для двухбуквенных блоков символов коды считаются точно так же. При этом $P((xy)) = P(x) \cdot P(y)$.

Код Шеннона.

Символ	Код
aa	10000
ab	0001
ac	10010
ad	10111
a<space>	11110100
a<EOF>	111111001
ba	0010
bb	0000
bc	0011
bd	01100
b<space>	11110001
b<EOF>	111110110
ca	10011
cb	0101
cc	10101
cd	11000
c<space>	11110110
c<EOF>	111111010
da	11010
db	01110
dc	11011
dd	11101
d<space>	111110010
d<EOF>	1111111011
<space>a	11110111
<space>b	11110011
<space>c	11111000
<space>d	111110100
<space><space>	11111111100
<space><EOF>	111111111100
<EOF>a	111111011
<EOF>b	111110111
<EOF>c	111111100
<EOF>d	1111111101
<EOF><space>	111111111101
<EOF><EOF>	1111111111110

Код Шеннона-Фано.

Символ	Код
aa	1000
ab	0010
ac	1001
ad	10111
a<space>	1111001
a<EOF>	11111100
ba	0011
bb	000
bc	0100
bd	0110
b<space>	11101
b<EOF>	11111010
ca	1010
cb	0101
cc	10110
cd	1100
c<space>	1111010
c<EOF>	111111010
da	11010
db	0111
dc	11011
dd	11100
d<space>	11111000
d<EOF>	111111101
<space>a	11110110
<space>b	1111000
<space>c	11110111
<space>d	11111001
<space><space>	111111110
<space><EOF>	1111111110
<EOF>a	111111011
<EOF>b	11111011
<EOF>c	111111100
<EOF>d	111111110
<EOF><space>	11111111110
<EOF><EOF>	11111111111

Код Хаффмана.

Символ	Код
aa	0001
ab	1100
ac	0100
ad	11101
a<space>	01111110
a<EOF>	011110101
ba	1101
bb	001
bc	1010
bd	1000
b<space>	0111010
b<EOF>	01110010
ca	0101
cb	1011
cc	0110
cd	11110
c<space>	01111111
c<EOF>	011110110
da	11111
db	1001
dc	0000
dd	11100
d<space>	01111001
d<EOF>	011100011
<space>a	01111100
<space>b	0111011
<space>c	01111101
<space>d	01111001
<space><space>	0111000100
<space><EOF>	011100010111
<EOF>a	011110111
<EOF>b	01110011
<EOF>c	01110000
<EOF>d	011110100
<EOF><space>	01110001010
<EOF><EOF>	011100010110

4) Код Шеннона:

010001000000000000001010001000000101001101100101001100000010100100000001011
00101100101101101101111010100010010010010010000010100010000001010000000101
10110001100101001000000110110010110010110100100111101001000100000100110010
00010100000001010000001101101000000010100100000011000110010110010010100111
1110

Код Шеннона-Фано:

10110111111111111001101111100100100110010011111100101111110001100011000100
10010001101101010101011110011011111001111110001001110011001011111001001100
01100010101000111011011110100110111001111110011111001001011111100101111100
111001100011010010000

Код Хаффмана:

11100110101010101101111010110100100111010011010110101101011001110011100100
10010001111001010101011011011110101101011001001100011101011010001001110
01110010101000110011110100100111101101101011011010001001011010110101101000
110001110011111010000

Код Шеннона (двухбуквенное кодирование):

000101010000000001001000010010110001101011000000010010010100101101011010111
01111110010000110101101010101100100001001001010010111010110010010010101100
11010110101101111110110001100010011110100010010100100101011001101100001001
0010101100011001011110000111111010

Код Шеннона-Фано (двухбуквенное кодирование):

001001010000000100100100011110011010110000010010101001111010110101110011111
00000101011010110010110010010001101010011111000110100101010110110101101011
01111110100100001001001101000110101001101010110110110001001010101100110101
111000111111010

Код Хаффмана (двухбуквенное кодирование):

11001011001001010011001101111101111111111000101001011110111111111111100011
11001110001100110101101001100110110111101111001000010010111000111111111100
00011111111010110010101111111011011110110111000000000101001011100010001110
10001011110110

5) Средняя длина кода для алфавита Λ вычисляется по формуле

$$L = \sum_{s \in \Lambda} P(s) \cdot L(s), \text{ где } L(s) - \text{длина кода символа } s.$$

Избыточность кода вычисляется по формуле $E = 1 - \frac{H}{L}$, где H - энтропия, L - средняя длина кода. В задаче 4 показано, считать энтропию. Подставляем числа и получаем:

Код	Средняя длина	Избыточность
Шеннона	2.777778	0.2332
Шеннона-Фано	2.25	0.053333
Хаффмана	2.25	0.053333
Шеннона (2х-букв)	4.806156	0.113318
Шеннона-Фано (2х-букв)	4.326818	0.015089
Хаффмана (2х-букв)	4.293467	0.007438

6) Напишем программу на языке Python, которая будет кодировать сообщение с использованием адаптивного сжатия по Хаффману. Результат:

```
101001111110100101110100100010001101011101001011111001011011011011011011
10101000110110010010001011101010111111001111110001001110011001011111001001
10001100010001000111010111110100101111001111110011111001001011111010110111
100111001010010110010000
```

7) Аналогично пункту 6:

```
00110000110110001010110001101111110100100011100011110111011001001100110011
10011011110100111111101001010010111010010000100000011110110100000010000000
11110100111101011101111010100001111101101011010110100001000100000111100101
0011010011110100111110110100111100000111111011110100101000100
```

8) Напишем программу на Python, которая будет кодировать сообщение с использованием арифметического кодирования. Так как точность числа с плавающей точкой сильно ограничена, будем считать все в дробях, а конечный результат переводить в десятичную дробь с высокой точностью с помощью сайта <https://matematika-club.ru/kalkulyator-bolshih-chisel>. Получим закодированное сообщение:

```
11011100111110100100010000111100010111000000100101100011101011101001100110
10000000110100111011100110011010011010011100010010111110000101100100001111
111000010110011000010001010101001111111111000000000100011111101111000100
1010011011000001001010110100100101011101100111101100110000001010000111111
000010110001111010111100010001
```

9) Результат кодирования сообщения алгоритмом LZW (с помощью Python):

```
00101000110010101010100001001101110010011010010000100100010110000011010011
00001001111001100100010100111101101111010000001110010101001100010010000010
01001101011000101001001000000101100101011000010100100000111100001101001110
00001000101001010001000101110011010101110001001001011010100000010011010001
10000000
```

Задача 5.

На языке программирования C# были реализованы два архиватора. Первый - на основе метода Хаффмана, второй - на основе алгоритма LZW. Код (с необходимыми комментариями) можно посмотреть в ветке FirstHW (Term 2/Homework 1/Archivers) или в pull request'e, соответствующем первой домашней работе.