МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ им. В. Г. ШУХОВА» (БГТУ им. В.Г. Шухова)

Кафедра программного обеспечения вычислительной техники и автоматизированных систем

Лабораторная работа №1

по дисциплине: «Теория информации»

Выполнил: ст. группы ПВ-211

Чувилко Илья Романович

Проверил:

Твердохлеб В.В.

Белгород 2023 г.

Тема работы: Исследование кодирования по методу Хаффмана. Оценка эффективности кода.

Цель работы: изучить способ кодирования сообщений по методу Хаффмана. Научиться строить дерево кода Хаффмана и составлять код Хаффмана по таблице вероятности появления символов в пределах алфавита исходного сообщения. Узнать, как вычисляется коэффициент сжатия, значение средней длины образовавшейся кодовой конструкции, величина дисперсии. Научиться выявлять наиболее эффективные из кодовых моделей сообщения по критериям коэффициента сжатия и дисперсии.

Выполнение работы:

Задание 1: Построить кодовое представление сообщения, вероятности появления символов в пределах алфавита которого приведены в табл.1.

Символ	s1	s2	s3	s4	s5	s6	s7	s8
Вероятность	0.23	0.19	0.16	0.16	0.10	0.10	0.05	0.01

Таблица 1 – Вероятности появления символов в пределах алфавита исходного сообщения

Задание 2. Построить кодовое представление сообщения, вероятности появления символов в пределах алфавита которого приведены в табл.2.

Символ	s1	s2	s3	s4	s5	s6	s7	s8
Вероятность	0.25	0.22	0.13	0.11	0.1	0.09	0.07	0.03

Таблица 2 – Вероятности появления символов в пределах алфавита исходного сообщения

Задание 3. Для условий, приведенных в заданиях 1 и 2, выявить возможность построения альтернативных кодовых моделей сообщения. В случае обнаружения таковых, выявить наиболее эффективные из них по критериям K_{comp} и δ.

Содержание отчета

- решения заданий 1-3 (процесс и результаты построения кодов Хаффмана, можно сжато, оценка полученных кодов и выводы касательно оптимального кода, если где-то будут возможны его разные варианты);
- скриншоты и краткое описание (пояснение процесса) исследования зависимости особенностей сообщений и результирующих кодов (приложение Хаффман);
- размещение сведений о работе с greedy – по желанию;
- выводы

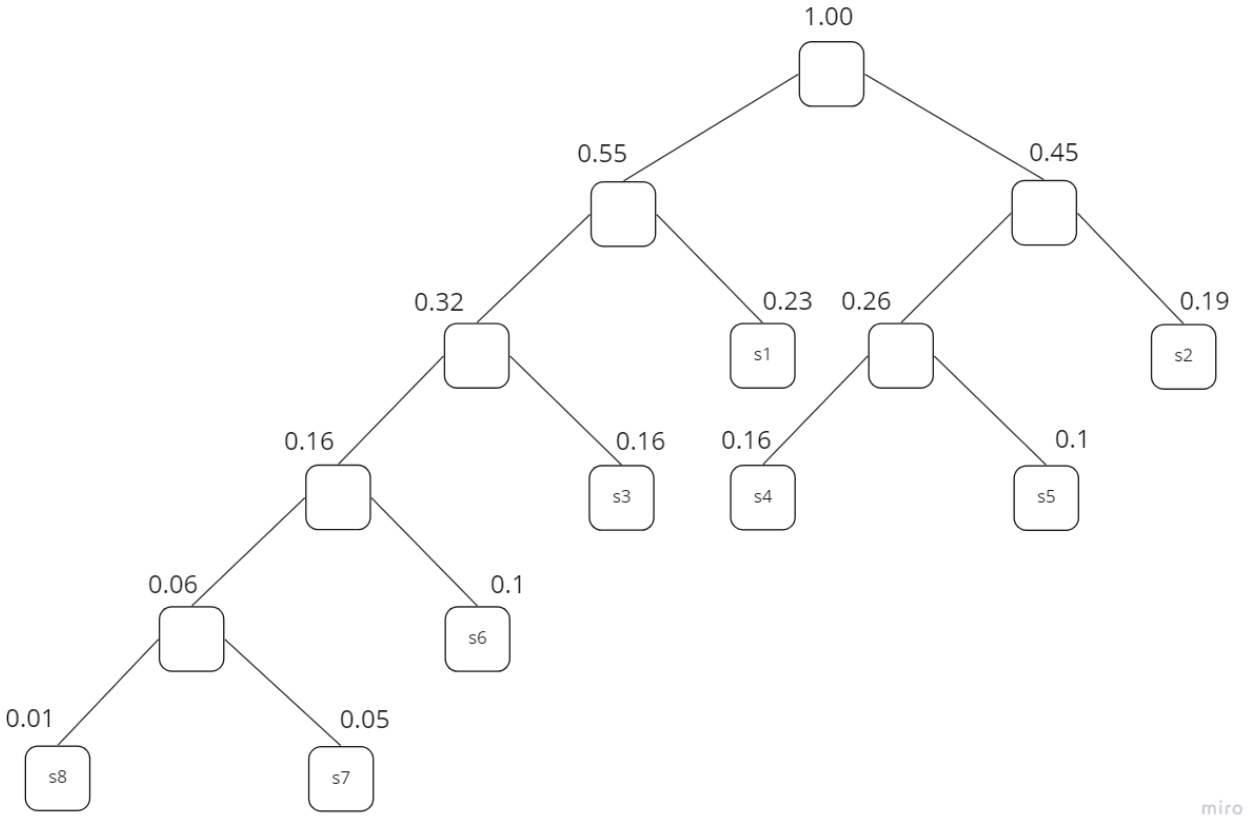
Выполнение работы:

Для получения кода по Хаффману воспользуемся методом деревьев по Хаффману.

1. В таблицах символов и их вероятностей каждая пара «символ-вероятность» рассматривается как один узел-лист дерева Хаффмана.
2. На первом шаге построения дерева выбирается пара узлов-листов, которым соответствуют минимальные величины вероятностей. Для них строится новый узел, вес которого будет равно сумме вероятностей входящих в него узлов-листов. Если на этом шаге, и на любом последующем, присутствует возможность выбора нескольких вариантов действий (например, есть более 2 узлов с минимальными и равными между собой вероятностями), можно выбирать любой из вариантов.
3. Новый узел, полученный на шаге 2, будет являться родительским по отношению к формирующим его узлам. С ними он соединяется ребрами, каждому из которых присваивается вес – 0 или 1 (можем назначать произвольно, но лучше в рамках текущего дерева придерживаться какого-то единого правила. Например – первым идет 1, или наоборот).
4. Пункты 2-3 повторяются до тех пор, пока не будет получена единая вершина (корень дерева), вес которой будет равен 1.
5. Двигаясь от корня по направлению к каждому узлу-листу, соответствующему тому или иному символу, считываются веса ребер. Из них формируется цепочка двоичных элементов, которая и будет кодом символа.

Задание 1

Построили дерево Хаффмана.



С его помощью получаем кодировку Хаффмана.

S1	10
S2	00
S3	110
S4	011
S5	010
S6	1110
S7	11110
S8	11111

Посчитаем параметры этого способа кодирования.

Для того, чтобы посчитать коэффициент сжатия допустим, что мы сгенерировали сообщение длинной в 100 символов с заданными вероятностями.

Его вес в 8 битной кодировке B = 8 * 100 = 800 бит.

Вес при кодировании полученным кодом $B' = \sum_i p_i * l_i = 280$

$$K = \frac{B}{B'} = \frac{800}{280} = 2,86$$

Вычислим дисперсию.

$$\ell_{ср} = \sum_i p_i \times \ell_i = 2,8 \; \text{наивп}$$

$$\delta = \sum_i p_i \times (\ell_i - \ell_{cp})^2$$

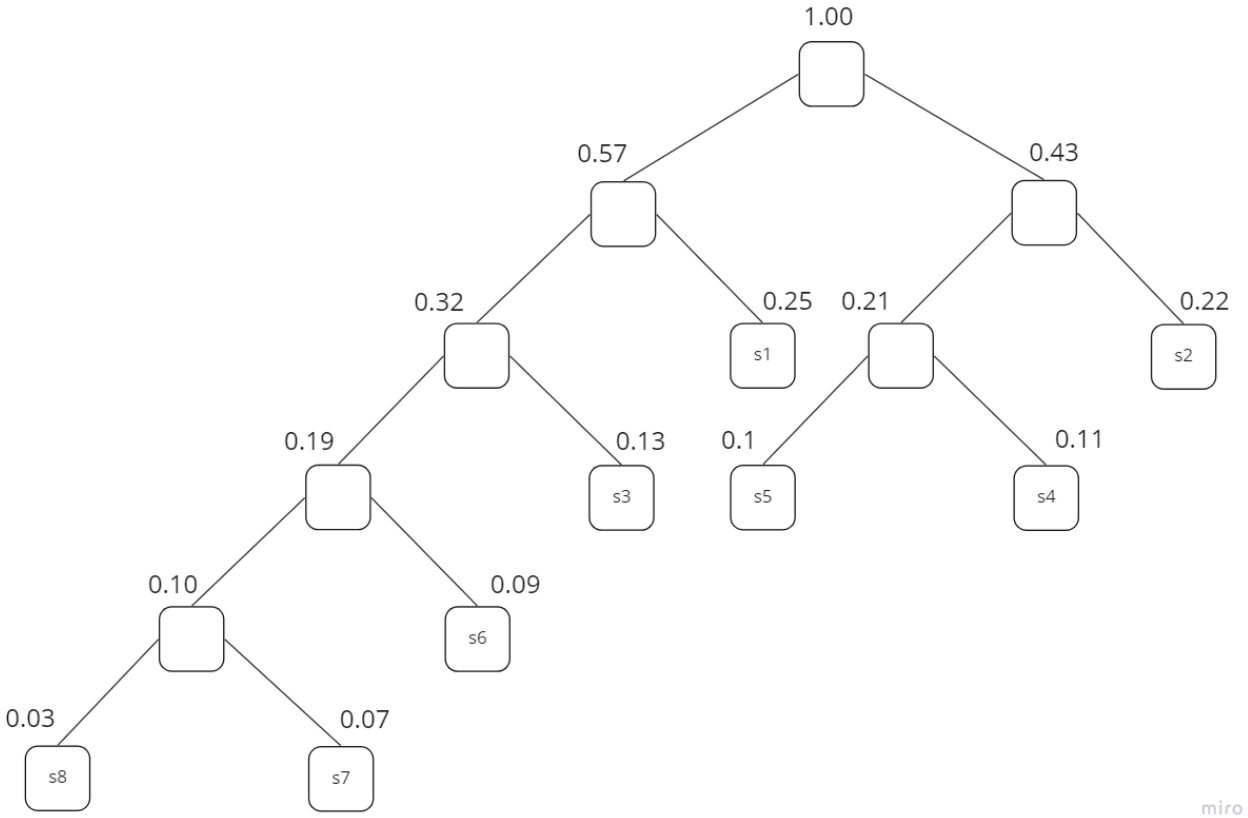
$\delta = 0,72$

В данном случае есть вариации дерева Хаффмана, но на эффективность кодирования они не влияют, поэтому и рассматривать их не нужно.

Задание 2

В данном случае есть возможность построить два вида деревьев, дающих различные коды и их характеристики.

Рассмотрим первое дерево.



С его помощью получаем кодировку Хаффмана.

S1	10
S2	00
S3	110
S4	010
S5	011
S6	1110
S7	11110
S8	11111

Посчитаем параметры этого способа кодирования.

Для того, чтобы посчитать коэффициент сжатия допустим, что мы сгенерировали сообщение

длинной в 100 символов с заданными вероятностями.
Его вес в 8 битной кодировке B = 8 * 100 = 800 бит.
Вес при кодировании полученным кодом $B^{'}=\sum_i p_i * l_i = 282$

$$K=\frac{B}{B'}=\frac{800}{282}=2{,}84$$

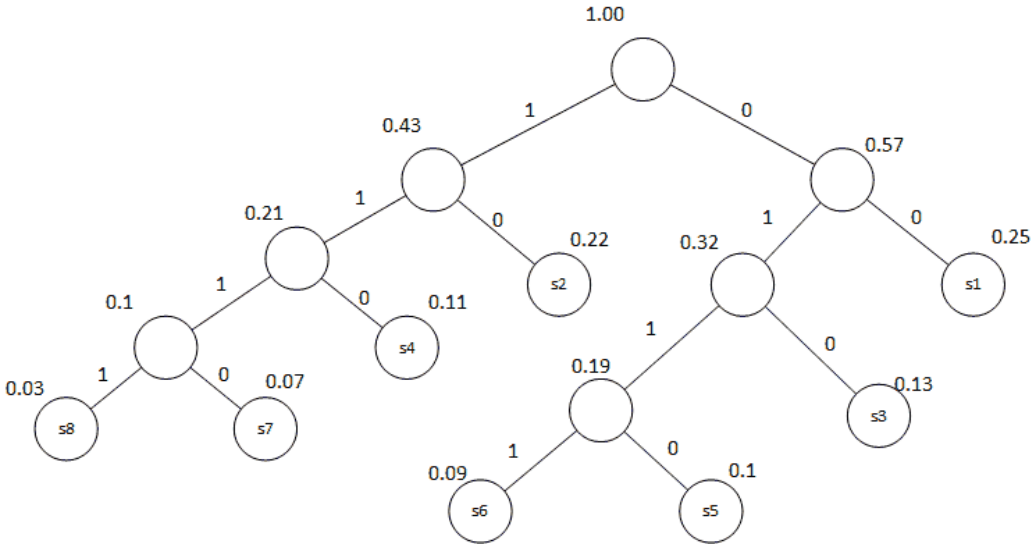
Вычислим дисперсию.

$$\ell_{cp}=\sum_i p_i \times \ell_i \quad {}_{=2},82$$

$$\delta=\sum_i p_i \times (\ell_i-\ell_{cp})^2$$

$$\delta=0{,}927\,6$$

Рассмотрим второе дерево.



С его помощью получаем кодировку Хаффмана.

S1	00
S2	10
S3	010
S4	110
S5	0110
S6	0111
S7	1110
S8	1111

Посчитаем параметры этого способа кодирования.
Для того, чтобы посчитать коэффициент сжатия допустим, что мы сгенерировали сообщение длинной в 100 символов с заданными вероятностями.
Его вес в 8 битной кодировке B = 8 * 100 = 800 бит.
Вес при кодировании полученным кодом $B^{'}=\sum_i p_i * l_i = 282$

$$K = \frac{B}{B'} = \frac{800}{282} = 2{,}84$$

Вычислим дисперсию.

$$\ell_{cp} = \sum_i p_i \times \ell_i = 2{,}82$$

$$\delta = \sum_i p_i \times (\ell_i - \ell_{cp})^2$$

$$\delta = 0{,}7276$$

Посчитаем параметры этого способа кодирования.

В данном случае вторая кодировка превосходит первую по показателю дисперсии, а по коэффициенту сжатия они одинаковы. Вторая кодировка более предпочтительна.

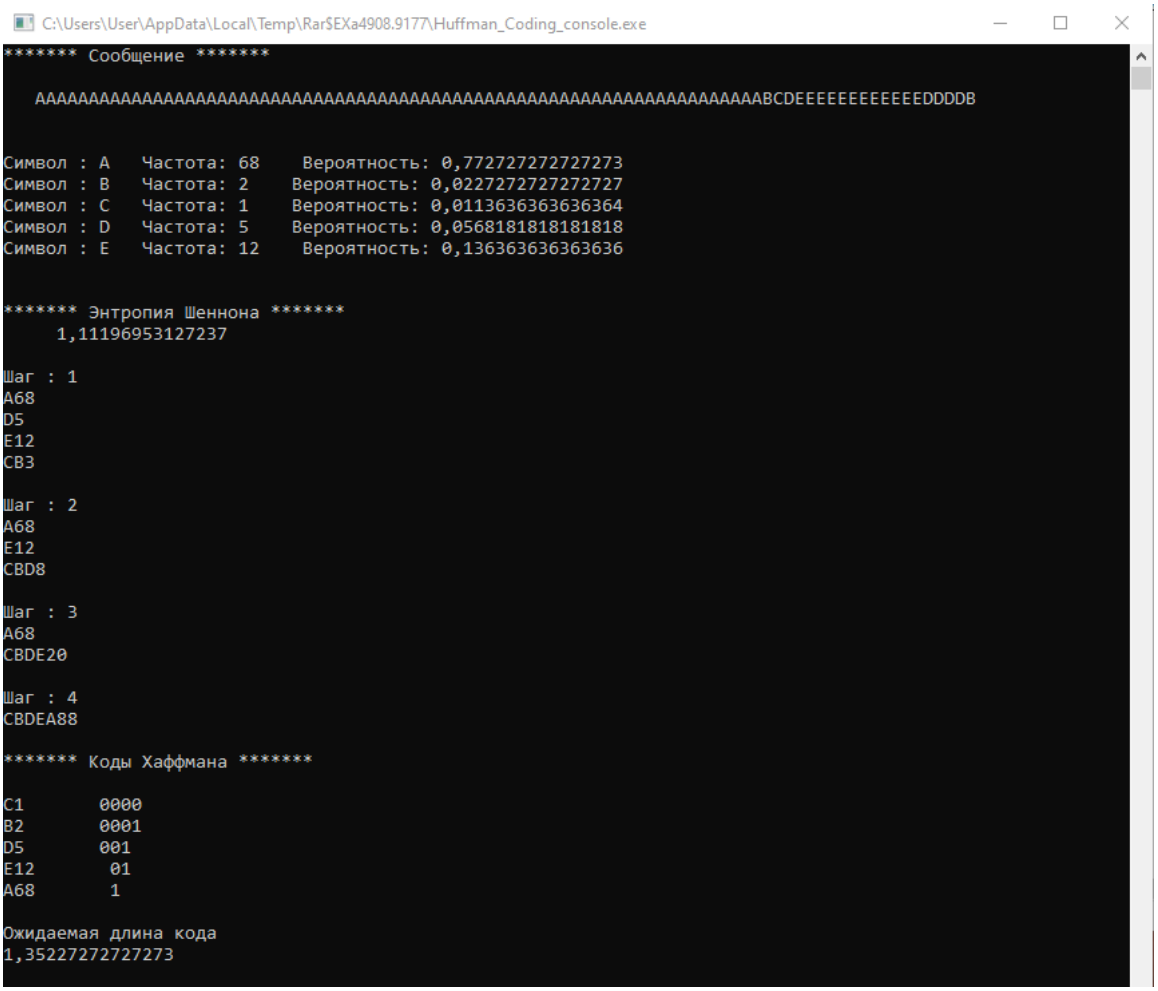
Задание 3

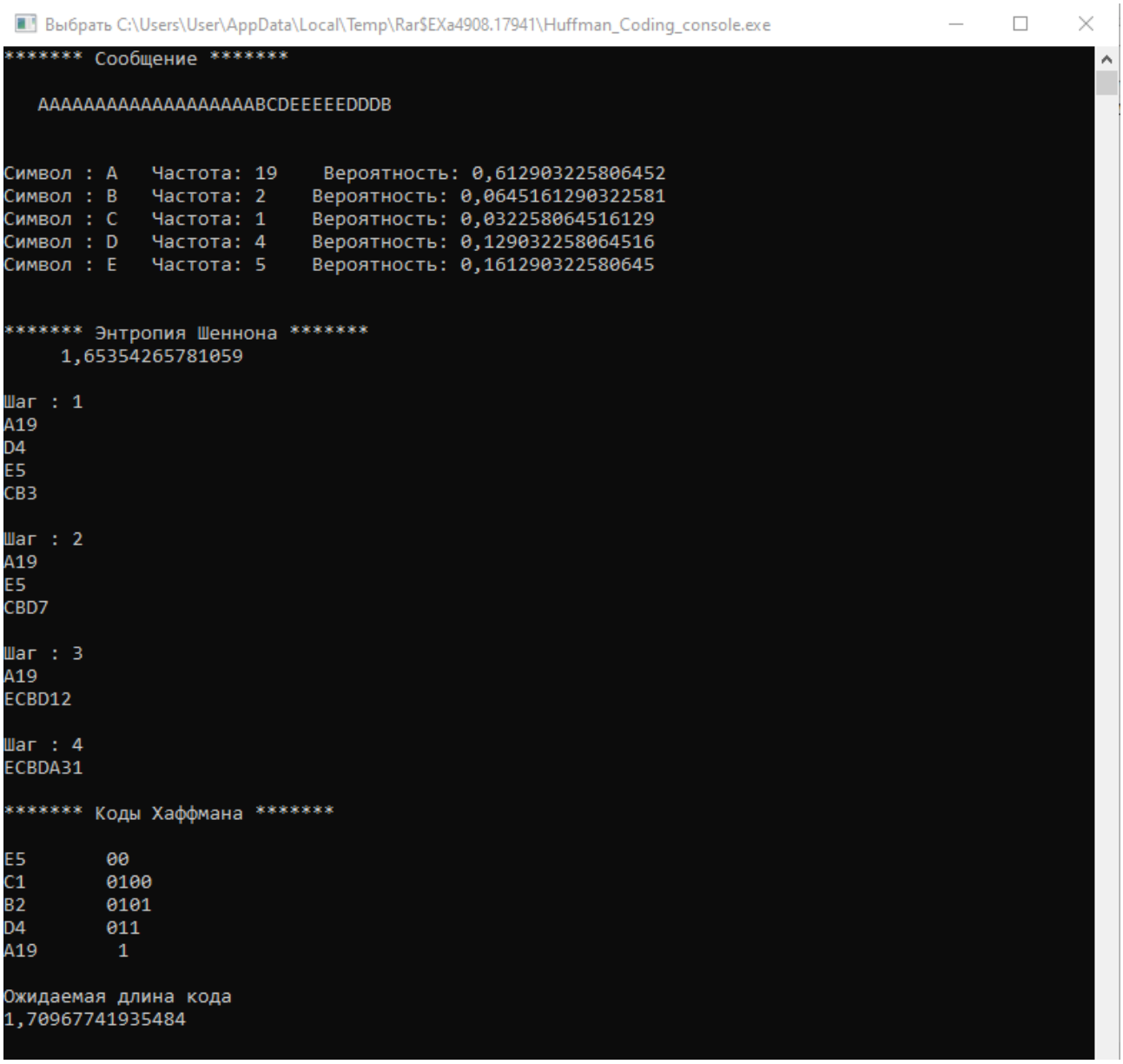
Выполнено в ходе рассмотрения заданий 1 и 2.

В ходе работы с программой Huffman_Coding_console, были замечены следующие зависимости.

Если в сообщение присутствуем много различных символов, энтропия Шеннона и средняя длина кода возрастают.

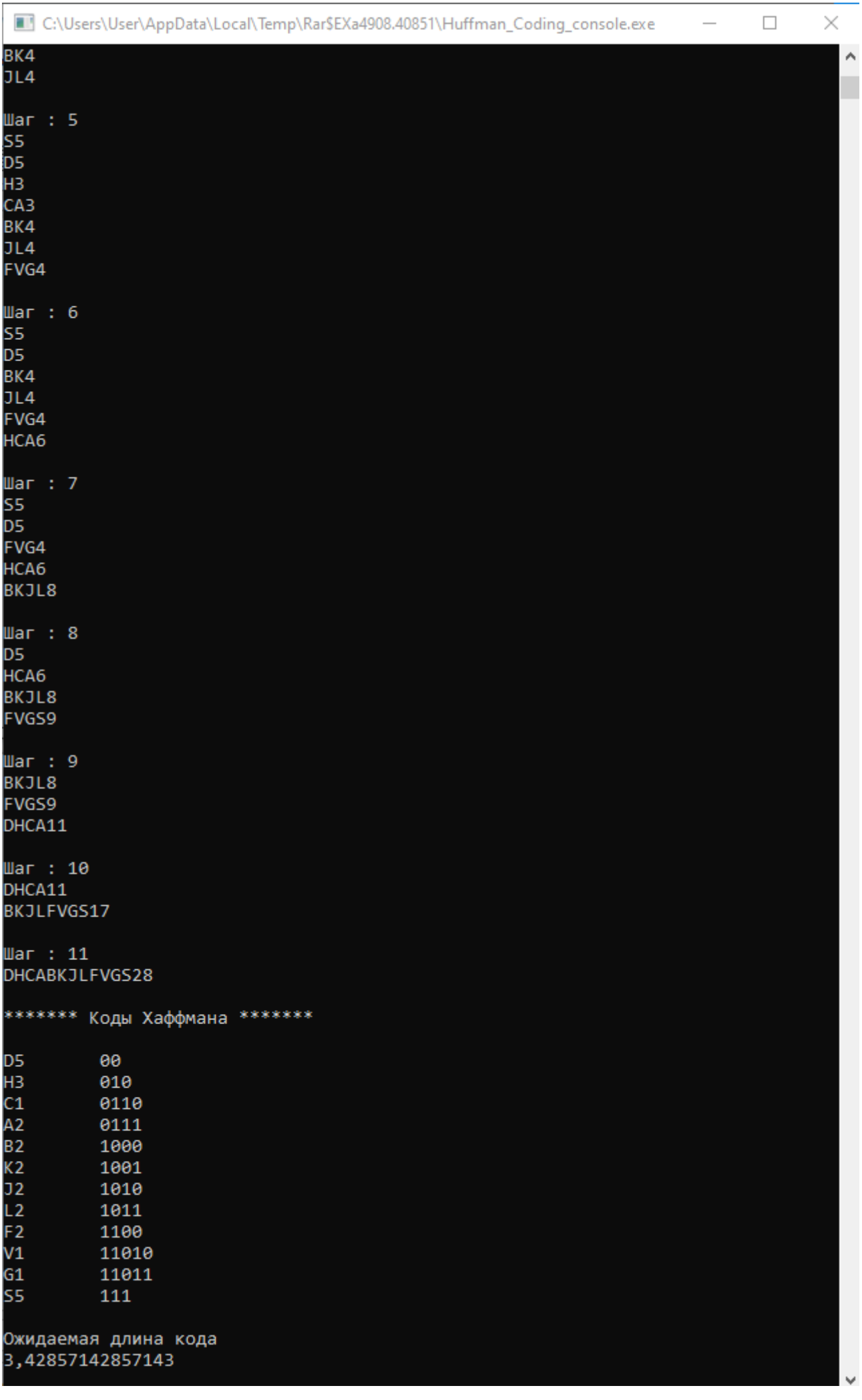
Если в сообщении будет большая доля символов лишь некоторых типов, то энтропия Шеннона и средняя длина кода принимают низкое значение. Коды Хаффмана хорошо сжимают сообщения, где много одинаковых символов.




Выбрать C:\Users\User\AppData\Local\Temp\Rar$EXa4908.17941\Huffman_Coding_console.exe
******* Сообщение *******
AAAAAAAAAAAAAAAAAAABCDEEEEEDDDB
Символ : A Частота: 19 Вероятность: 0,612903225806452
Символ : B Частота: 2 Вероятность: 0,0645161290322581
Символ : C Частота: 1 Вероятность: 0,032258064516129
Символ : D Частота: 4 Вероятность: 0,129032258064516
Символ : E Частота: 5 Вероятность: 0,161290322580645
******* Энтропия Шеннона *******
1,65354265781059
Шаг : 1
A19
D4
E5
CB3
Шаг : 2
A19
E5
CBD7
Шаг : 3
A19
ECBD12
Шаг : 4
ECBDA31
******* Коды Хаффмана *******
E5 00
C1 0100
B2 0101
D4 011
A19 1
Ожидаемая длина кода
1,70967741935484

C:\Users\User\AppData\Local\Temp\Rar$EXa4908.36554\Huffman_Coding_console....

```
******* Сообщение *******

   SDHAVDSGHDSBAKJSLLDFSDFKHJCB


Символ : S   Частота: 5    Вероятность: 0,178571428571429
Символ : D   Частота: 5    Вероятность: 0,178571428571429
Символ : H   Частота: 3    Вероятность: 0,107142857142857
Символ : A   Частота: 2    Вероятность: 0,0714285714285714
Символ : V   Частота: 1    Вероятность: 0,0357142857142857
Символ : G   Частота: 1    Вероятность: 0,0357142857142857
Символ : B   Частота: 2    Вероятность: 0,0714285714285714
Символ : K   Частота: 2    Вероятность: 0,0714285714285714
Символ : J   Частота: 2    Вероятность: 0,0714285714285714
Символ : L   Частота: 2    Вероятность: 0,0714285714285714
Символ : F   Частота: 2    Вероятность: 0,0714285714285714
Символ : C   Частота: 1    Вероятность: 0,0357142857142857


******* Энтропия Шеннона *******
     3,37970604880628

Шаг : 1
S5
D5
H3
A2
B2
K2
J2
L2
F2
C1
VG2

Шаг : 2
S5
D5
H3
B2
K2
J2
L2
F2
VG2
CA3

Шаг : 3
S5
D5
H3
J2
L2
F2
VG2
CA3
BK4

Шаг : 4
S5
D5
H3
F2
```


C:\Users\User\AppData\Local\Temp\Rar$EXa4908.40851\Huffman_Coding_console.exe
BK4
JL4
War : 5
S5
D5
H3
CA3
BK4
JL4
FVG4
War : 6
S5
D5
BK4
JL4
FVG4
HCA6
War : 7
S5
D5
FVG4
HCA6
BKJL8
War : 8
D5
HCA6
BKJL8
FVGS9
War : 9
BKJL8
FVGS9
DHCA11
War : 10
DHCA11
BKJLFVGS17
War : 11
DHCABKJLFVGS28
******* Коды Хаффмана *******
D5 00
H3 010
C1 0110
A2 0111
B2 1000
K2 1001
J2 1010
L2 1011
F2 1100
V1 11010
G1 11011
S5 111
Ожидаемая длина кода
3,42857142857143

Вывод: в ходе работы изучен способ кодирования сообщений по методу Хаффмана. Получены навыки составления кода Хаффмана по таблице вероятности появления символов в пределах алфавита исходного сообщения, нахождения коэффициента сжатия, значения средней длины образовавшейся кодовой конструкции, величины дисперсии, выявления наиболее эффективных из кодовых моделей сообщения по критериям коэффициента сжатия и дисперсии.