Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Институт информатики и вычислительной техники

09.03.01 "Информатика и вычислительная техника"
профиль "Программное обеспечение средств вычислительной техники и автоматизированных систем"

Практическая работа №2

по дисциплине «Теория информации»

«Побуквенное кодирование текстов»

Выполнил: студент 4 курса
ИВТ, гр. ИП-111
Кузьменок Д.В.

Работу проверил: доцент кафедры ПМиК
Мачикина Елена Павловна

Новосибирск 2025

Цель работы:

Экспериментальное изучение избыточности сжатия текстового файла.

Задание:

1. Запрограммировать процедуру двоичного кодирования текстового файла побуквенным кодом. В качестве методов сжатия использовать метод Хаффмана и метод Шеннона (или метод Фано). Текстовые файлы использовать те же, что и в практической работе 1.
2. Вычислить среднюю длину кодовых слов и оценить избыточность кодирования для каждого построенного побуквенного кода.
3. После кодирования текстового файла вычислить оценки энтропии файла с закодированным текстом H_1, H_2, H_3 (после кодирования последовательность содержит 0 и 1) и заполнить таблицу.

Результаты работы

```
Processing file1.txt:
Huffman codes:
Symbol: A, Code: 0
Symbol: B, Code: 10
Symbol: C, Code: 11
ShennonFano codes:
Symbol: A, Code: 00
Symbol: C, Code: 01
Symbol: B, Code: 1
For 1-symbol groups:
Huffman: Entropy = 0,9709154507296759, Average Code Length = 1,665, Redundancy = 0,6940845492703241
Shannon-Fano: Entropy = 0,9700178420868771, Average Code Length = 1,6691, Redundancy = 0,6990821579131229
```

```
Huffman codes:
Symbol: BA, Code: 000
Symbol: AC, Code: 001
Symbol: CB, Code: 010
Symbol: AA, Code: 011
Symbol: AB, Code: 100
Symbol: CA, Code: 101
Symbol: BC, Code: 110
Symbol: BB, Code: 1110
Symbol: CC, Code: 1111
ShennonFano codes:
Symbol: BC, Code: 0000
Symbol: CA, Code: 0001
Symbol: AB, Code: 001
Symbol: AA, Code: 010
Symbol: CB, Code: 011
Symbol: AC, Code: 100
Symbol: BA, Code: 101
Symbol: CC, Code: 110
Symbol: BB, Code: 111
```

Генерация кодов для сиволов

```
Processing file1.txt:
Huffman codes:
ShennonFano codes:
For 1-symbol groups:
Huffman: Entropy = 0,9709154507296759, Average Code Length = 1,665, Redundancy = 0,6940845492703241
Shannon-Fano: Entropy = 0,9700178420868771, Average Code Length = 1,6691, Redundancy = 0,6990821579131229

Huffman codes:
ShennonFano codes:
For 2-symbol groups:
Huffman: Entropy = 0,9925213460226432, Average Code Length = 3,2136213621362133, Redundancy = 2,22110001611357
Shannon-Fano: Entropy = 0,9897407063536849, Average Code Length = 3,231623162316232, Redundancy = 2,241882455962547

Huffman codes:
ShennonFano codes:
For 3-symbol groups:
Huffman: Entropy = 0,9959678102852264, Average Code Length = 4,799559911982396, Redundancy = 3,8035921016971694
Shannon-Fano: Entropy = 0,9955211798976311, Average Code Length = 4,818363672734546, Redundancy = 3,822842492836915
```

Результаты для первого файла

```
Processing file2.txt:
Huffman codes:
ShennonFano codes:
For 1-symbol groups:
Huffman: Entropy = 0,9964592533353147, Average Code Length = 1,4935999999999998, Redundancy = 0,4971407466646851
Shannon-Fano: Entropy = 0,9964592533353147, Average Code Length = 1,4935999999999998, Redundancy = 0,4971407466646851

Huffman codes:
ShennonFano codes:
For 2-symbol groups:
Huffman: Entropy = 0,9934641456135784, Average Code Length = 2,987198719871987, Redundancy = 1,9937345742584087
Shannon-Fano: Entropy = 0,9538660308976981, Average Code Length = 3,1732173217321735, Redundancy = 2,2193512908344752

Huffman codes:
ShennonFano codes:
For 3-symbol groups:
Huffman: Entropy = 0,9961890185333067, Average Code Length = 4,46869373874775, Redundancy = 3,4725047202144435
Shannon-Fano: Entropy = 0,9910768471119484, Average Code Length = 4,544308861772356, Redundancy = 3,5532320146604075
```

Результаты для второго файла

```
Processing file3.txt:
Huffman codes:
ShennonFano codes:
For 1-symbol groups:
Huffman: Entropy = 0,9962613506897541, Average Code Length = 4,383371077321257, Redundancy = 3,387109726631503
Shannon-Fano: Entropy = 0,9837736678251969, Average Code Length = 4,507387037636151, Redundancy = 3,5236133698109535

Huffman codes:
ShennonFano codes:
For 2-symbol groups:
Huffman: Entropy = 0,9985302909444176, Average Code Length = 7,986626402070725, Redundancy = 6,988096111126308
Shannon-Fano: Entropy = 0,9975234375573316, Average Code Length = 8,033110440034486, Redundancy = 7,035587002477154

Huffman codes:
ShennonFano codes:
For 3-symbol groups:
Huffman: Entropy = 0,9993129030678473, Average Code Length = 10,660338690539612, Redundancy = 9,661025787471765
Shannon-Fano: Entropy = 0,9995124140003921, Average Code Length = 10,696149282708895, Redundancy = 9,696636868708502
```

Результаты для третьего файла

Метод кодирования	Название текста	Оценка избыточности кодирования	H_1	H_2	H_3
Код Хаффмана	Три символа с одинаковыми вероятностями	0,6940845492703241	0,9709154507296759	0,9925213460226432	0,9959678102852264
Код Фано	Три символа с одинаковыми вероятностями	0,6990821579131229	0,9700178420868771	0,9897407063536849	0,9955211798976311
Код Хаффмана	Три символа с разными вероятностями (0.5, 0.2, 0.3)	0,4971407466646851	0,9964592533353147	0,9934641456135784	0,9961890185333067
Код Фано	Три символа с разными вероятностями (0.5, 0.2, 0.3)	0,4971407466646851	0,9964592533353147	0,9934641456135784	0,9961890185333067
Код Хаффмана	1984 – Джордж Оруэлл (английский текст)	3,387109726631503	0,9962613506897541	0,9985302909444176	0,9993129030678473
Код Фано	1984 – Джордж Оруэлл (английский текст)	3,5236133698109535	0,9837736678251969	0,9975234375573316	0,9995124140003921

Вывод:

При кодированиях были получены префиксные код, в которых используется избыточность сообщения (коды более частых символов состоят из коротких последовательностей, а коды более редких символов – из более длинных).

Можно увидеть, что данные методы кодирования обладают высокой избыточностью. Энтропия полученных последовательностей близка к единице, что говорит о том, что на один символ приходится один бит информации. При том, при выборе пар или троек символов энтропия почти не меняется. Это говорит о том, что символы в получившихся кодах равновероятны, что подтверждает эффективность кодирования.