Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Сибирский государственный университет телекоммуникаций и
информатики»
(СибГУТИ)

Кафедра прикладной математики и кибернетики

Теория Информации

Лабораторная работа №4

Выполнил:
Студент IV курса ИВТ,
группы ИП-713
Михеев Никита Алексеевич

Работу проверил:
доцент кафедры ПМиК
Мачикина Е. П.

Новосибирск 2020 г.

1. Постановка задачи

Цель работы:

Экспериментальное изучение процесса сжатия текстового файла с помощью бинарного кодирования.

Задание:

1. Запрограммировать процедуру двоичного кодирования текстового файла. В качестве метода кодирования использовать или метод Шеннона, или метод Фано, или метод Хаффмана. Текстовые файлы использовать те же, что и в практических работах 1, 2, 3.

2. Проверить, что построенный код для каждого файла является префиксным. Вычислить среднюю длину кодового слова и оценить избыточность каждого построенного кода.

3. После кодирования текстового файла вычислить оценки энтропии выходной последовательности, используя частоты отдельных символов, пар символов и троек символов и заполнить таблицу.

4. Оформить отчет, загрузить отчет и файл с исходным кодом в электронную среду. Отчет обязательно должен содержать заполненную таблицу и анализ полученных результатов.

2. Ход работы

Для выполнения лабораторной работы была написана программа на языке Python версии 3.9, которая сначала считывает текстовый файл с исходным кодом, приводит его к удобному для работы программы виду. Для кодирования текста был выбран метод Шеннона.

Затем идет подсчет каждого символа (или их последовательности) и вычисляются вероятности их встречи относительно всего текста. Далее вычисляются длины кодовых слов, а также кумулятивные вероятности.

После этого в программе начинается вычисление кодов символов: берется двоичное представление кумулятивной вероятности встречи символа, из дробной части берется такое количество символов, сколько было подсчитано выше. Результат выводится на экран.

3. Результат работы

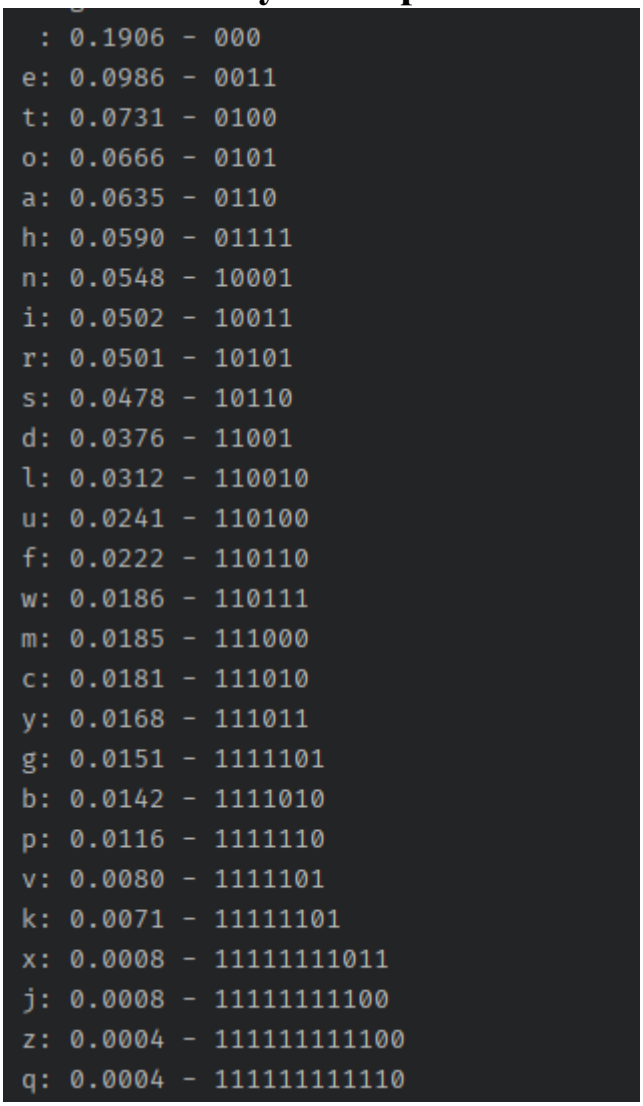

Рис.1 – вывод на экран символа, его вероятности и кодовое слово.

Рис.2 – энтропия источника

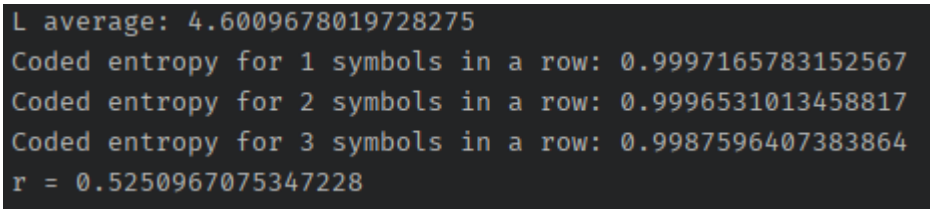

Рис.3 – средняя длина кодового слова, энтропия закодированного текста, избыточность кодирования.

Метод кодирования	Название текста	Оценка избыточности кодирования	Оценка энтропии выходной посл-ти (частоты символов)	Оценка энтропии выходной посл-ти (частоты пар символов)	Оценка энтропии выходной посл-ти (частоты троек символов)
Метод Шеннона	«The White Company»	0.5251	0.99971	0.99965	0.99876

Таблица 1 – полученные данные

Вывод:

При кодировании методом Шеннона был получен префиксный код, в котором используется избыточность сообщения (коды более частых символов состоят из коротких последовательностей, а коды более редких символов – из более длинных). По нашим данным выяснилось, что энтропия оригинального текста далека от реальных значений и поэтому избыточность получилась сравнительно небольшой. Но, если взять энтропию реально текста (около 2), то избыточность будет большой (~2.6), что показывает не оптимальность кодировки Шеннона.