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

Кафедра прикладной математики и кибернетики

Теория Информации

Лабораторная работа №2

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 свойств энтропии Шеннона для текстов на естественном языке

Задание:

1. Выбрать художественный текст на русском (английском) языке. Объем файла в формате txt более 10 Кб. Для алфавита текста предполагается, что строчные и заглавные символы не отличаются, знаки препинания опущены, к алфавиту добавлен пробел, для русских текстов буквы «е» и «ё», «ь» и «ъ» совпадают.

2. Составить программу, определяющую несколько оценок энтропии данного текстового файла. Оценки энтропии необходимо вычислить по формуле Шеннона двумя способами, т.е. используя частоты отдельных символов и используя частоты пар символов. По желанию можно продолжить процесс вычисления оценок с использованием частот троек, четверок символов и т.д.

3. После тестирования программы необходимо заполнить таблицу для отчета и проанализировать полученные результаты. Сравнить полученные результаты с результатами работы 1.

4. Оформить отчет, загрузить отчет и файл с исходным кодом в электронную среду. Отчет обязательно должен содержать заполненную таблицу и анализ полученных результатов. По желанию в отчет можно включить описание программной реализации. В отчет не нужно включать содержимое этого файла.

2. Ход работы

Для выполнения лабораторной работы была написана программа на языке Python версии 3.9, которая сначала считывает текстовый файл, приводит его к удобному для работы программы виду – избавляется от лишних символов, оставляя только символы пробела и английского алфавита.

Затем считаный файл в программе делится на последовательности по 1 и 2 символа и идет вычисление отношения этих последовательностей к общему количеству, так же подсчитывается энтропия.

3. Результат работы

```
"C:\Users\Lolimpo\Google Drive\SibSUTIS\Labs\4 course\Inf. Theory\Labs\venv\Scripts\python.exe"
 "C:/Users/Lolimpo/Google Drive/SibSUTIS/Labs/4 course/Inf. Theory/Labs/lab2.py"
For 1 symbols in a row: Shanon entropy: 4.075871094438105
For 2 symbols in a row: Shanon entropy: 3.504620051985727

Process finished with exit code 0
```

Рис.1 – результат работы программы

Название текста	Максимально возможное значение энтропии	Оценка энтропии (одиночные символы)	Оценка энтропии (частоты пар символов)
The White Company by Sir Arthur Conan Doyle	4.7549	4.075	3.505

Таблица 1 – полученные данные

Выводы:

1. При одиночных символах, из-за различных вероятностей встретить тот или иной символ, энтропия не достигает максимальной;
2. При парных символах, энтропия снижается, так как не все цепочки встречаются из возможных в тексте и в художественном тексте существует множество правил и определенностей.