

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

Кафедра прикладной математики и кибернетики

Теория Информации
Лабораторная работа №3

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. Для выполнения работы используйте один из файлов с исходным кодом программы, который был написан для практических работ 1, 2 или 3. Определите алфавит исходного кода, текст комментариев и ввода/вывода на экран игнорировать.

2. Составить программу, определяющую несколько оценок энтропии файла с исходным кодом. 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 и 2.

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 с исходным кодом, приводит его к удобному для работы программы виду.

Затем считанный 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/lab3.py"  
For 1 symbols in a row: Shanon entropy: 4.996551712137453  
For 2 symbols in a row: Shanon entropy: 3.1176716364117687  
  
Process finished with exit code 0
```

Рис.1 – результат работы программы

Язык программирования	Максимально возможное значение энтропии	Оценка энтропии (одиночные символы)	Оценка энтропии (частоты пар символов)
Python	5.555	4.9966	3.1177

Таблица 1 – полученные данные

Вывод:

Энтропия ниже, так как в языках программирования имеется большее количество строгих грамматических правил, чем в естественных языках. Для одиночных символов энтропия завышена, так как учитываются символы пунктуации, не входящие в естественный язык.