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

Кафедра прикладной математики и кибернетики

Теория Информации

Лабораторная работа №1

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 при рассмотрении текстов с заданным алфавитом.

Задание:

1. Для выполнения работы необходимо предварительно сгенерировать два файла. Каждый файл содержит последовательность символов, количество различных символов больше 2 (3,4 или 5). Объем файлов больше 10 Кб, формат txt. Символы последовательно и независимо генерируются с помощью датчика псевдослучайных чисел и записываются в файл. Первый файл (назовем его F1) должен содержать последовательность символов с равномерным распределением, т.е. символы должны порождаться равновероятно и независимо. Для генерации второго файла (F2) необходимо сначала задать набор вероятностей символов, а затем последовательно и независимо генерировать символы с соответствующей вероятностью и записывать их в файл.

2. Составить программу, определяющую несколько оценок энтропии созданных текстовых файлов. Вычисление значения по формуле Шеннона настоятельно рекомендуется оформить в виде отдельной функции, на вход которой подается массив (список) вероятностей, выходной параметр – значение, вычисленное по формуле Шеннона.

3. После тестирования программы необходимо заполнить таблицу для отчета и проанализировать полученные результаты. Для получения теоретических значений энтропии используйте наборы вероятностей, которые использовались при генерации файлов.

4. Оформить отчет, загрузить отчет и файл с исходным кодом в электронную среду.

2. Ход работы

Для выполнения лабораторной работы была написана программа на языке Python версии 3.9, которая сначала генерирует два файла с символьными последовательностями из 50000 символов, с алфавитом из 4 символов (a, b, c, d) с определенными вероятностями. В файле F1 вероятности равные по 0.25, в файле F2 разные – 0.5, 0.2, 0.2 и 0.1 соответственно для каждого символа алфавита.

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

3. Результат работы

```
Alphabet: {'a': 0.25, 'b': 0.25, 'c': 0.25, 'd': 0.25}
Probabilities: [('a', 0.2478), ('b', 0.25028), ('c', 0.25234), ('d', 0.24958)]
For 1 symbols in a row: Shanon entropy: 1.9999695079899964
Probabilities: [('aa', 0.05962), ('ab', 0.06262), ('ac', 0.06158), ('ad', 0.06326), ('ba', 0.06226),
 ('bb', 0.06358), ('bc', 0.06522), ('bd', 0.06418), ('ca', 0.06454), ('cb', 0.05938), ('cc', 0.0643),
 ('cd', 0.06358), ('da', 0.06194), ('db', 0.05958), ('dc', 0.06162), ('dd', 0.06242)]
For 2 symbols in a row: Shanon entropy: 1.9993062086661955

Alphabet: {'a': 0.5, 'b': 0.2, 'c': 0.2, 'd': 0.1}
Probabilities: [('a', 0.50054), ('b', 0.19856), ('c', 0.2014), ('d', 0.0995)]
For 1 symbols in a row: Shanon entropy: 1.7597334296381795
Probabilities: [('aa', 0.25234), ('ab', 0.10034), ('ac', 0.09958), ('ad', 0.04758), ('ba', 0.09906),
 ('bb', 0.0381), ('bc', 0.0411), ('bd', 0.02038), ('ca', 0.09934), ('cb', 0.04018), ('cc', 0.04134),
 ('cd', 0.0197), ('da', 0.05034), ('db', 0.0197), ('dc', 0.02006), ('dd', 0.01054)]
For 2 symbols in a row: Shanon entropy: 1.7591298560858308


Process finished with exit code 0
```

Рис.1 – результат работы программы

	Оценка энтропии (частоты отдельных символов	Теоретическое значение энтропии (отдельные символы)	Оценка энтропии (частоты пар символов)	Теоретическое значение энтропии (для пар символов)
F1	1.9999	2	1.9993	2
F2	1.7597	1.761	1.7591	1.761

Таблица 1 – полученные данные

Выводы:

1. При равновероятном распределении достигается максимальное значение энтропии;
2. При независимой генерации символов, энтропия остается неизменной при любой длине последовательности символов.