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 №1

по дисциплине «Теория информации»

«Вычисление энтропии Шеннона»

Выполнил: студент 4 курса

ИВТ, гр. ИП-111

Кузьменок Д.В.

Работу проверил: доцент кафедры ПМиК

Мачикина Елена Павловна

Новосибирск 2025

**Цель работы:**

Экспериментальное вычисление оценок энтропии Шеннона текстов. Изучение свойств энтропии Шеннона.

**Задание:**

1. Для выполнения работы потребуются три текстовых файла с различными свойствами. Объем файлов больше 10 Кб, формат txt. В первом файле содержится последовательность символов, количество различных символов больше 2 (3,4 или 5). Символы **последовательно и независимо** с равными вероятностями генерируются с помощью датчика псевдослучайных чисел и записываются в файл.

Для генерации второго файла необходимо сначала задать набор вероятностей символов (количество символов такое же, как и в первом файле), а затем **последовательно и независимо** генерировать символы с соответствующей вероятностью и записывать их в файл, вероятности в процессе записи файла не меняются.

В качестве третьего файла необходимо выбрать художественный текст на русском (английском) языке. 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.

1. 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 символов, выходной параметр – значение, вычисленное по формуле Шеннона.

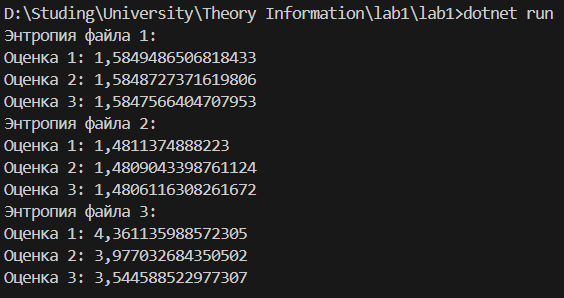
Вычислить три оценки энтропии Шеннона для каждого из файлов. Рекомендуется вычисление оценки оформить в виде отдельной функции с параметром имя файла:

**Первая оценка **. Сначала определить частоты отдельных символов файла, т.е. отношения количества отдельного символа к общему количеству символов в файле. Далее используя полученные частоты как оценки вероятностей, рассчитать оценку энтропии по формуле Шеннона.

**Вторая оценка **. Определить частоты всех последовательных пар символов в файле. Для того правильной оценки энтропии **** пары символов нужно рассматривать с перехлестом.

**Третья оценка **. Определить частоты всех последовательных троек символов в файле. Для того правильной оценки энтропии ** тройки** символов нужно рассматривать с перехлестом.

**Результаты работы**



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Название файла |  |  |  | Максимально возможное значение энтропии | Теоретическое значение энтропии |
| файл 1 | 1.584949 | 1.584873 | 1.584757 | 1.585 | 1.585 |
| файл 2 | 1.481137 | 1.480904 | 1.480612 | 1.585 | 1.485 |
| файл 3 | 4.361136 | 3.977033 | 3.544589 |  |  |

Максимальное и теоретические значения энтропии для файла 1 вычисляется с помощью формулы Хартли (т.к. вероятности символы равны между друг другом) как , где N – количество символов алфавита:

≈ 1.585.

Максимальное значение энтропии для файла 2 вычисляется с помощью формулы Хартли (т.к. энтропия текста при разных вероятностях символов не может превышать энтропию равновероятных символов) как , где N – количество символов алфавита:

≈ 1.585.

Теоретическое значение Энтропии вычисляется с помощью формулы Шеннона:

H = -Ʃ = -(0.5 + 0.3 + 0.2 = 1.485



Теоретическое значение энтропии (для пар символов):

1. Для файла 1:

H2(0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111; 0.1111) = (0.3522138890491458 \* 9) / 2 ≈1.5848732.

2) Для файла 2:

AA – 0,5 \* 0,5 = 0,25

AB – 0,5 \* 0,3 = 0,15

AC – 0,5 \* 0,2 = 0,1

BA – 0,3 \* 0,5 = 0,15

BB – 0,3 \* 0,3 = 0,09

BC – 0,3 \* 0,2 = 0,06

CA – 0,2 \* 0,5 = 0,1

CB – 0,2 \* 0,3 = 0,06

CC – 0,2 \* 0,2 = 0,04

H2 (0,25; 0,15; 0,1; 0,15; 0,09; 0,06; 0,1; 0,06; 0,04) = 2.97095/ 2 ≈1.485475

**Вывод:**

Сравнив теоретические и практические значения энтропии, можно сказать, что они очень близки. Из этого можно сделать вывод, что программа работает верно.