Университет ИТМО

Факультет ПИиКТ

Прикладная математика

Лабораторная работа №1

«Вычисление энтропии Шеннона»

Нестеров Дали Константинович

Группа P3302

**Цель работы:**

Получить практические навыки решения задач на количественное измерение информационного объема текстовой информации.

**Задание:**

1. Реализовать процедуру вычисления энтропии для текстового файла. Требования к текстовому файлу:

• документ на английском языке

• размер текстового файла - 20-60кБ

• документ должен быть осмысленным

В процедуре необходимо подсчитывать частоты появления символов (прописные и заглавные буквы не отличаются, знаки препинания рассматриваются как один символ, пробел является самостоятельным символом), которые можно использовать как оценки вероятностей появления символов. Затем вычислить величину энтропии. Точность вычисления -- 4 знака после запятой. Обязательно предусмотреть возможность ввода имени файла, для которого будет вычисляться энтропия.

2. Проверить запрограммированную процедуру на нескольких файлах и заполнить таблицу 1.1. вычисленными значениями энтропии

3. Вычислить значение энтропии для тех же файлов, но с использованием частот вхождений пар символов и заполнить таблицу 1.2

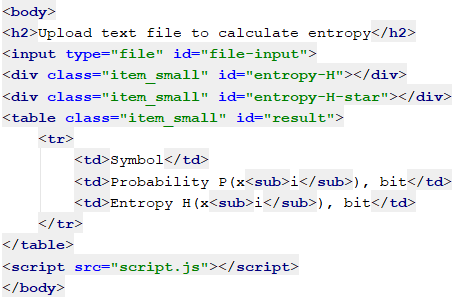
4. Проанализировать полученные результаты.

**Описание входных данных:**

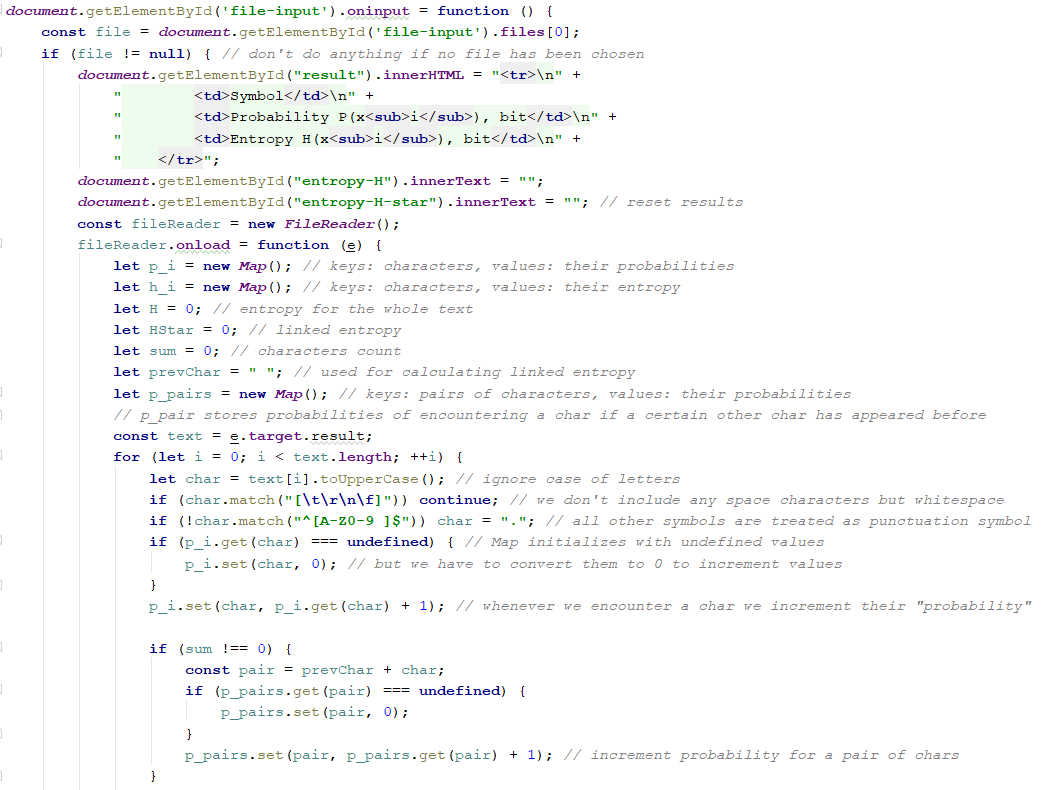
В качестве входных данных были выбраны отрывки из трех произведений на английском языке: «Nightfall», «Infinite jest» и «Evgeny Onegin»

**Решение поставленной задачи:**

Index.html(body):

****

Script.js():





**Результат работы программы:**

Таблица 1.2. для файла nightfall\_demo.txt

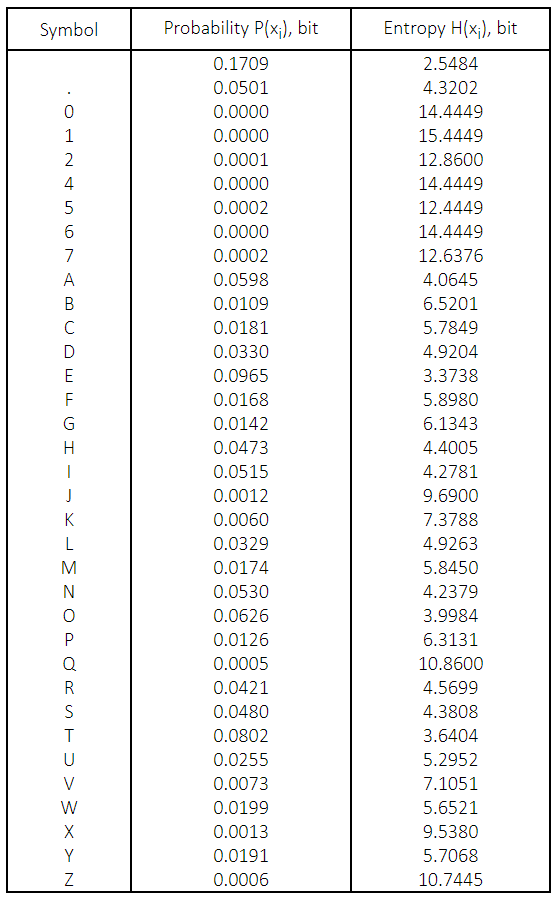


Таблица 1.2.

|  |  |  |  |
| --- | --- | --- | --- |
| Файл: | nightfall\_demo.txt | infinite\_jest\_demo.txt | evgeny\_onegin\_demo.txt |
| Энтропия H(X), бит | 4.1837 | 4.2148 | 3.9986 |
| Энтропия H\*(X), бит | 0.5123 | 0.5103 | 0.6284 |

**Вывод:** в ходе лабораторной работы была реализована программа для расчета энтропии и были получены практические навыки решения задач на количественное измерение информационного объема текстовой информации. Энтропия везде была примерно одинаковой. Энтропия с учетом пар символов получается значительно меньше, чем без и условие 0 <= H\*(X) <= H(X) соблюдается.