Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут ім. Ігоря Сікорського”

Фізико-технічний інститут

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

з предмету «Криптографія»

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконав:

Студент 3 курсу,

ФТІ, групи ФБ-05

Савченко Ярослав

Київ - 2022

**Мета роботи**

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

**Порядок виконання роботи**

0. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

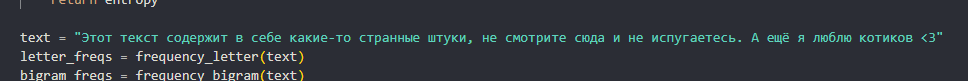
2. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30) H .

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

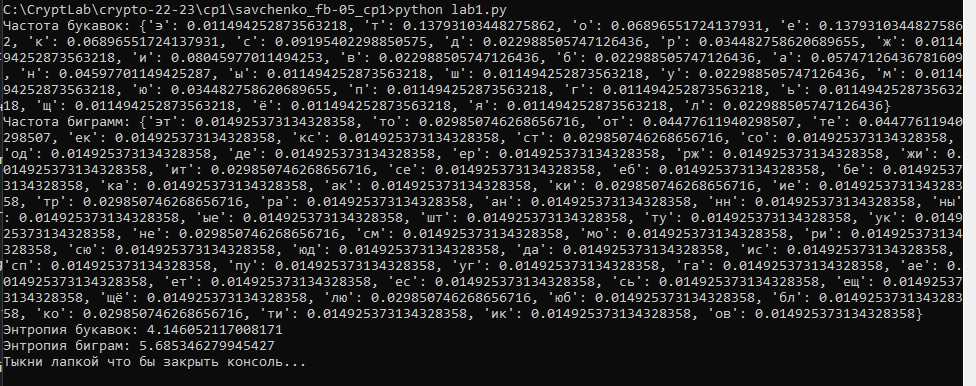
**Хід роботи:**

Напишемо наш прототип програми для виконання завдання 1

Файл lab1\_1.py містить в собі 3 функції для знаходження частоти літер, біграм та для знаходження ентропії. Пропишемо якийсь довільний текст для перевірки



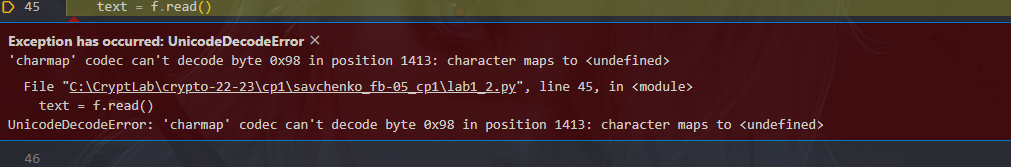
Перевіримо



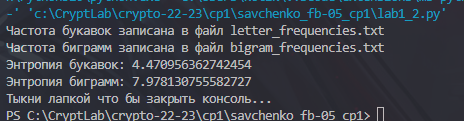
Чудово. Задачу виконав, але трохи модифікуємо програму. Додамо можливість записати результати частот у файл, для легшої роботи з даними. А також додамо можливість брати текст з файлу.

Модифікована версія знаходиться в файлі lab1\_2.py

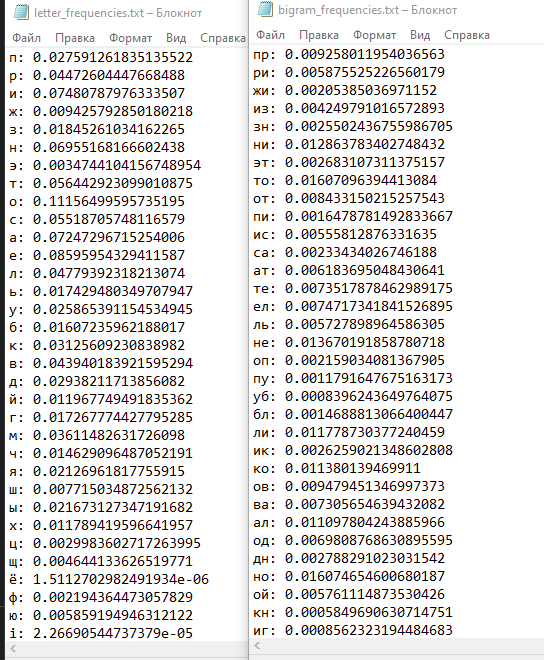
Тестуємо. В текстовий файл я додав видання мого олюбленого письменника Лавкрафта що розмір був більше 1 МБ



Перша проблема. Але вирішується простим ігноруванням символів які не можуть розпізнатися Unicode. Адже нам такі не потрібні.



Чудово, програма виконала свою задачу, подивимося на файли



Знову модифікуємо код, щоб відсортувати результат по спаданню.

Найбільш частими буквами та біграмами були:

|  |  |
| --- | --- |
| о: 0.11156499595735195  е: 0.08595954329411587  и: 0.07480787976333507  а: 0.07247296715254006  н: 0.06955168166602438  т: 0.056442923099010875  с: 0.05518705748116579  л: 0.04779392318213074  р: 0.04472604447668488  в: 0.043940183921595294 | ст: 0.017174470252385547  но: 0.016074654600680187  то: 0.01607096394413084  не: 0.013670191858780718  на: 0.013544709536102924  по: 0.01345059779409458  ни: 0.012863783402748432  ен: 0.01206475625981484  ра: 0.011872842119248803  ли: 0.011778730377240459 |

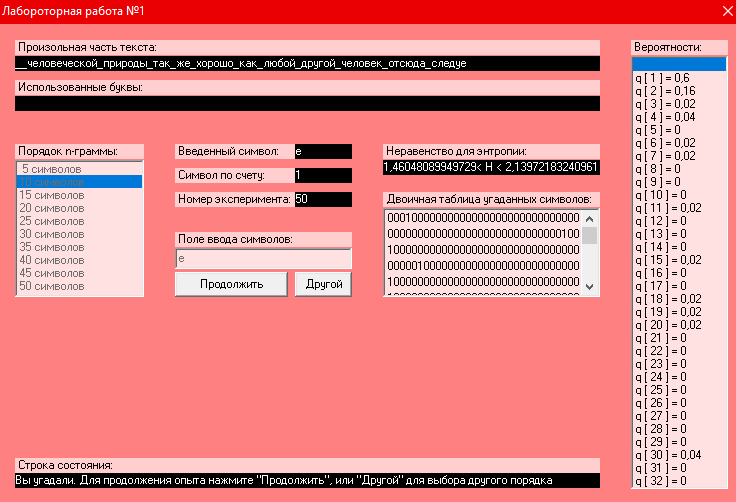
Ентропія:

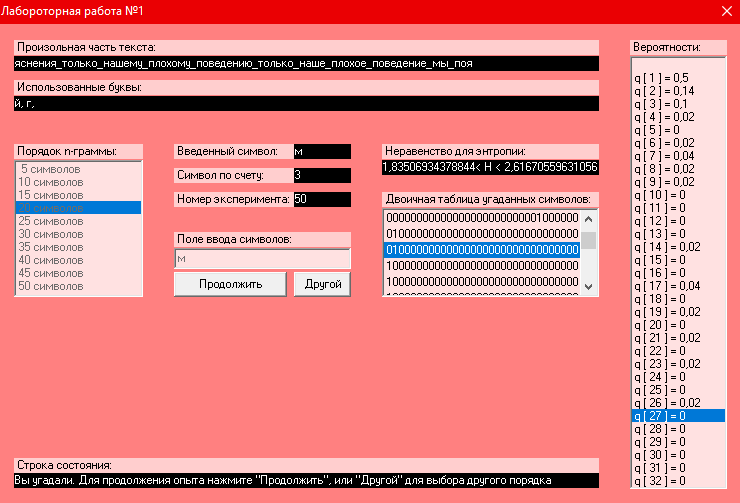
|  |  |
| --- | --- |
| 4.470956362742454 | 7.978130755582727 |

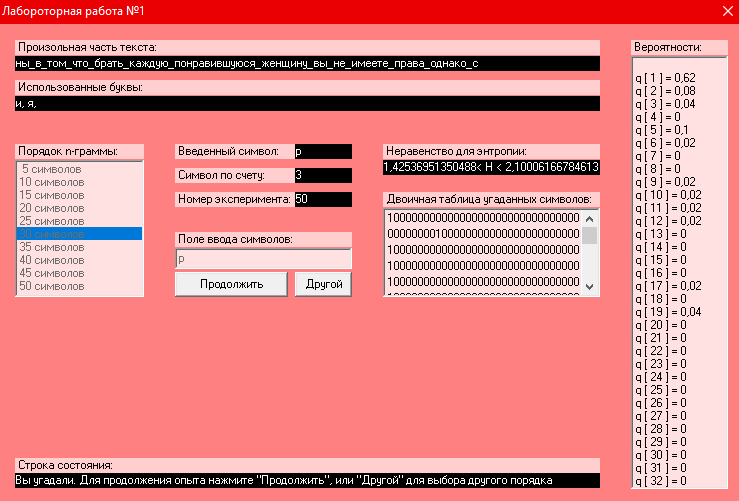
Надлишковість:

|  |  |
| --- | --- |
| 0,1058087275 | 0,2021869244 |

Перейдемо до другої задачі. Запустимо програму CoolPinkProgram

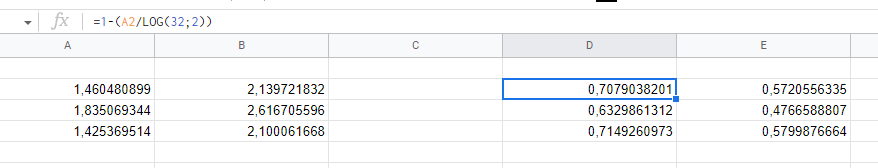






Тепер порахуємо оцінку надлишковості для кожного експерименту

Для цього я записав дані енропії в exel таблицю та функцією порахував надлишковість



Отже результат:

0,5720556335 < R < 0,7079038201

0,4766588807 < R < 0,6329861312

0,5799876664 < R < 0,7149260973

Висновки: на цій лабораторній роботі я отримав навички оцінки ентропії на символ джерела, навчився визначати наближене значення ентропії. А саме головне, це поборов помилку в якій використав не те кодування, ~~але таку безглузду проблему я не покажу~~ :)