Лабораторна робота 1

ФБ-11 Яцентюк Андрій ФБ-11 Кустов Іван (Варіант 4)

**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.

В цій лабі було вирахувано частоти монограм та біграм для тексту “Библия”.(Text.txt)

Текстові файли з частотами знаходяться в файлах results1.txt (Текст з пробілами)

results2.txt (Текст без пробілів).

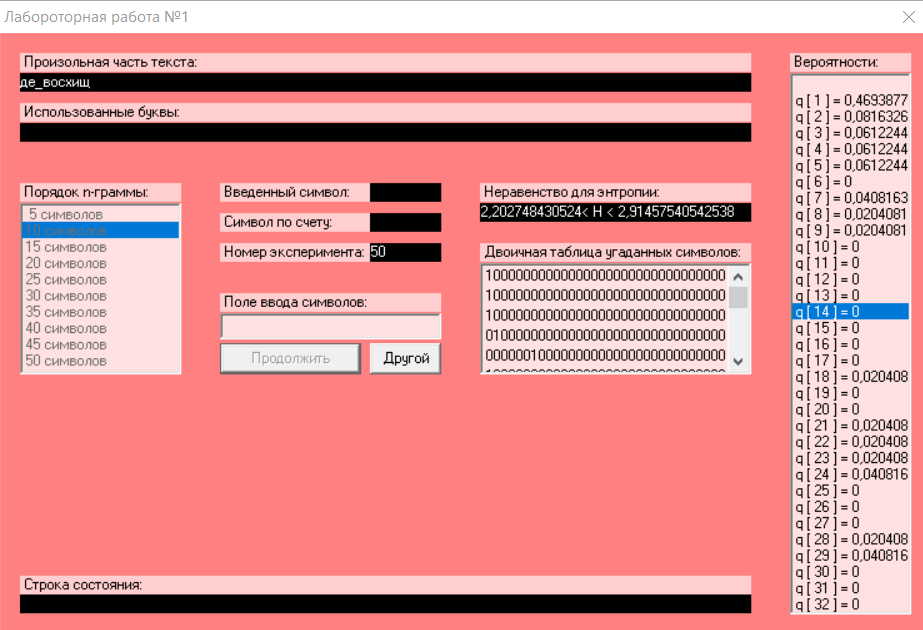
Частоти були вираховані за допомогою python скрипта, з використанням функції “**count**”.

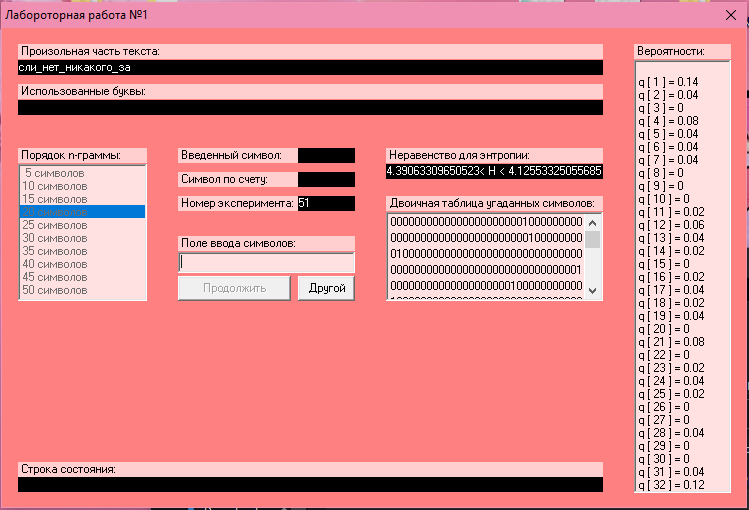
Ентропії були пораховані з використаням бібліотеки “**math**”

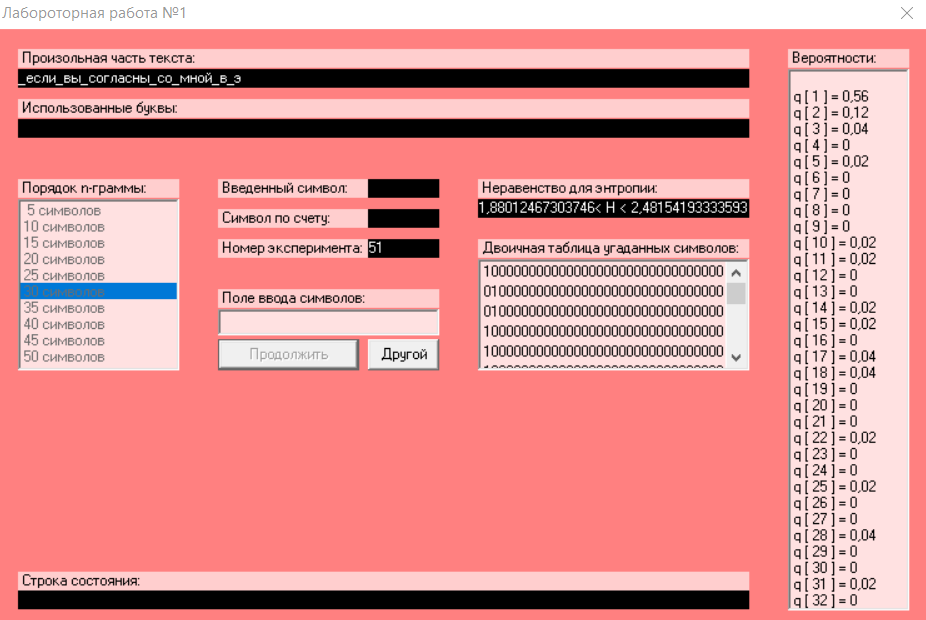
Вцілому робота була не дуже складною, через зручність використання мови програмування **Python**.

Найскладнішою частиною було розібратись з архітектурою **git** та **github.**

**Результат роботи**







Таблиці частот біграм знаходяться в файлах **bigram\_table\_russian1.xlsx(з пробілами) bigram\_table\_russian2.xlsx (без пробілів)**

**Частоти для монограм однакові, з пробілами, чи без.**

Взагалом, ентропія з CoolPinkProgram виявилась у середньому меншою за ту що підрахував скрипт.

**З пробілами**

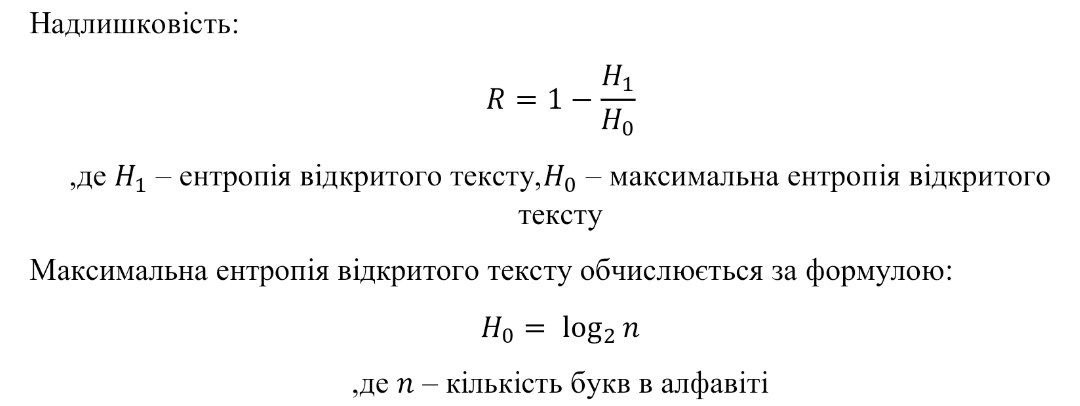
**Ентропія монограм: 4.433878685419471**

**Ентропія біграм: 3.930350462829067**

**Без пробілів**

**Ентропія монограм: 4.433878685419471**

**Ентропія біграм: 4.1144731556708365**

****

**H0 = log2(n) = log2(32) = 5**

**H1(З пробілами) = 4.909277396497351**

**H1(Без пробілів) = 4.433878685419471**

**H2(З пробілами)(З перетином) = 3.9261968370593516**

**H2(З пробілами)(Без перетинів) = 3.8500985125798097**

**H2(Без пробілами)(З перетином) = 4.1144731556708365**

**H2(Без пробілами)(Без перетинів) = 4.018163043724456**

**R1(З пробілами) = 0.181445207005298**

**R1(Без пробілів) = 0.1832242629161058**

**R2(З пробілами)(З перетином) = 0.2147603258812968**

**R2(З пробілами)(Без перетинів) = 0.22998029748403806**

**R2(Без пробілами)(З перетином) =** **0.1771053688658327**

**R2(Без пробілами)(Без перетинів) =**  **0.1963673912551088**

Вцілому, російська мова є досить передбачуваною якщо знати велику бібліотеку слів, та знати якусь кількість букв до необхідного місця. ***Крім того моменту коли останній відомий символ є пробілом. Тоді біда. Тоді вгадати першу букву слова є значно складнішою задачею.***

Частоти монограм майже співпадають з нормою наданою Вікіпедією.