**Міністерство освіти і науки України**

**Національний технічний університет України**

**"Київський політехнічний інститут імені Ігоря Сікорського"**

**Фізико-технічний інститут**

**Криптографія**

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Студенти ФБ-01

Новак О. І.

Тостоган Є. Г.

Київ 2022

**Мета роботи:** засвоєння поняття 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.

**Постановка задачі:**

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. За допомогою програми CoolPinkProgram оцінити значення H(10) , H(20) , H(30) . Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

**Хід роботи**

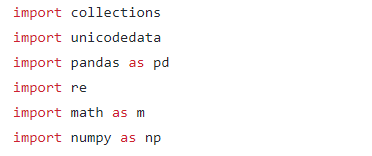
Для початку для визначення частоти входження символів в текст було написано програму на мові python3.

**Програма містить в собі такий функціонал:**

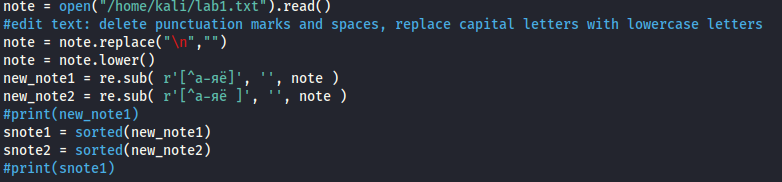
* Обробка тексту(видалення зайвих символів та абзаців, заміна регістру літер на малі)
* Підрахунок кількості входжень кожного символі в текст та відповідно визначення частоти
* Пошук біграм, що перетинаються та не перетинаються, та визначення частоти їх входжень у текст
* Обчислення ентропії Н1с та Н2 для тексту з пробілами та без
* Внесення результатів у датафрейми та таблиці excel.

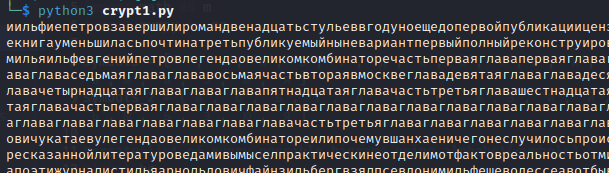
Код програми міститься у файлі crypto1.py

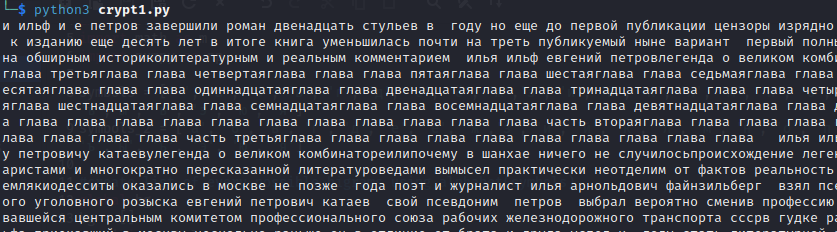
В ході роботи ми імпортували потрібні для реалізації программного коду бібліотеки:



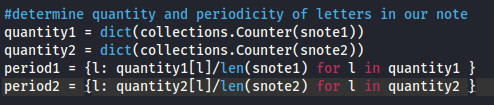
та провели обробку обраного тексту



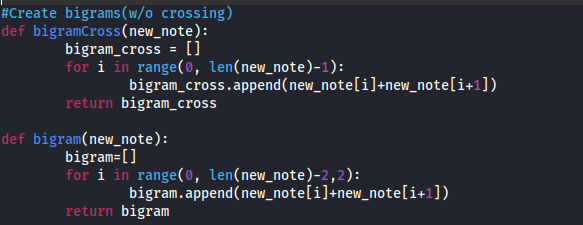




Рахуємо кількість та частоту появи літери в тексті:



Функції для створення біграм з утворених списків літер:



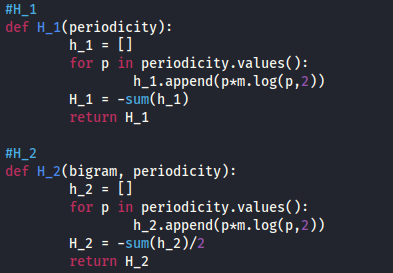
За допомогою отриманих функції створюємо біграми та перетворюємо їх у словники для подальшої роботи з ними.



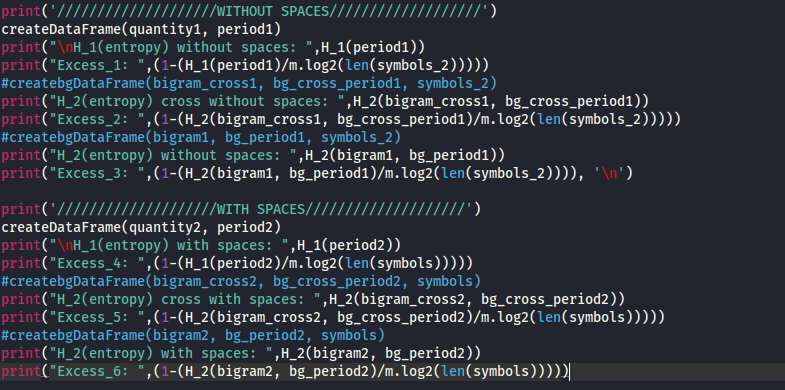
Функції для створення дата фреймів та таблиць для літер та біграм. У ході виконання стикнулися з проблемою перезапису таблиці, тому імя таблиць задаємо окремо з командного рядка.

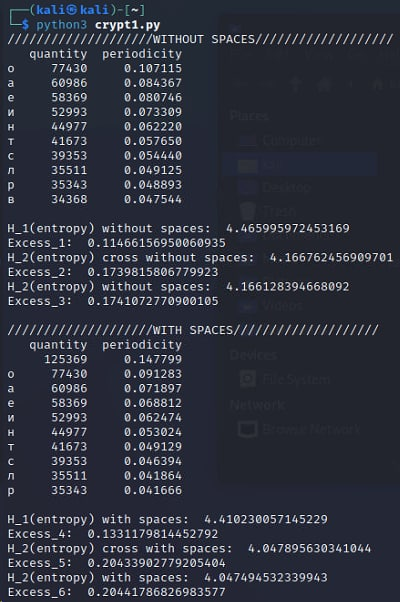


Функції для підрахунку ентропії з отриманих даних:



Вивід результатів:



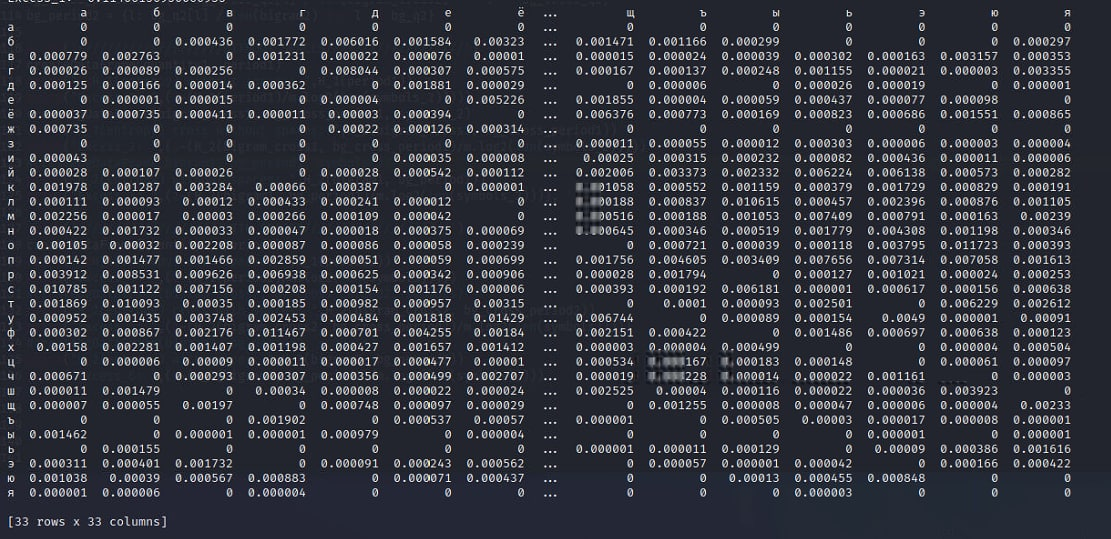


**Ентропія**

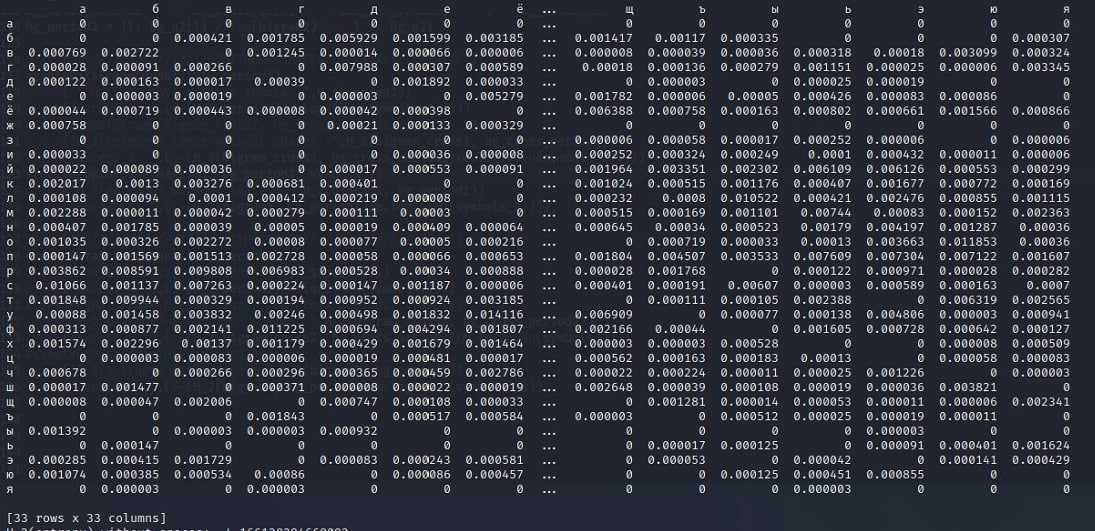
|  |  |  |
| --- | --- | --- |
|  | Текст з пробілами | Текст без пробілів |
| Н1 | 4.410230057145229 | 4.465995972453169 |
| Н2 | 4.047895630341044 | 4.166762456909701 |
| Н2(біграми з кроком 2) | 4.047494532339943 | 4.166128394668092 |
|  |  |  |

Для утворення матриці частот попередньо було додано два алфавіти у вигляді списку. Матриці частот з отриманих даних:

Bigram\_cross1()



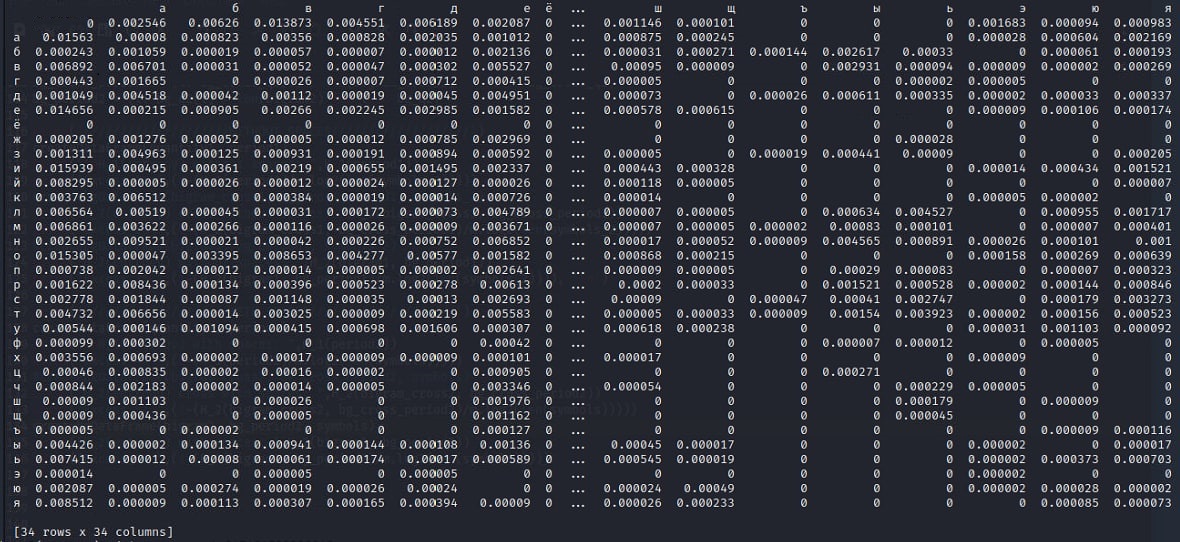
Bigram1()



Bigram\_cross2()



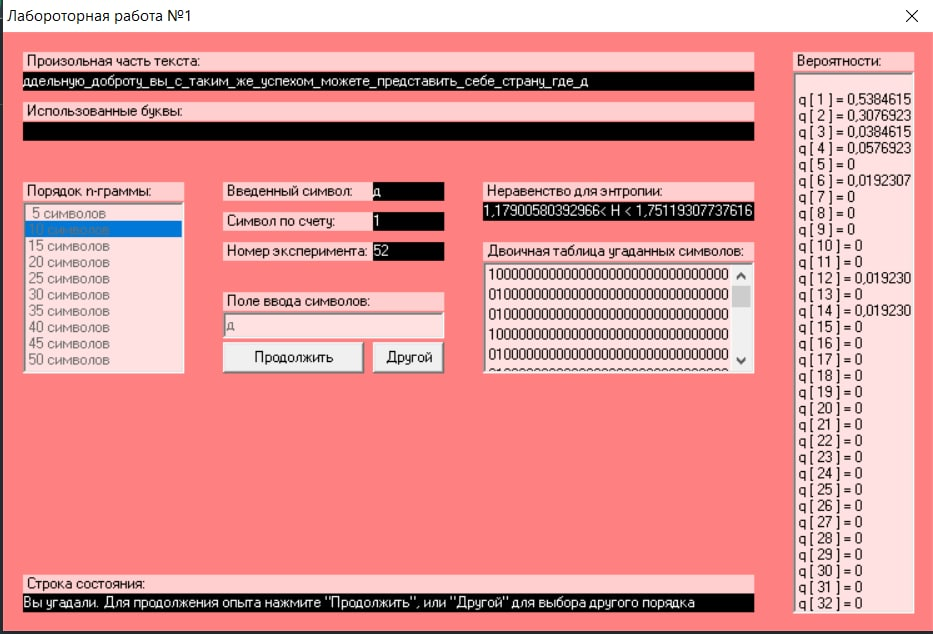
Bigram2()



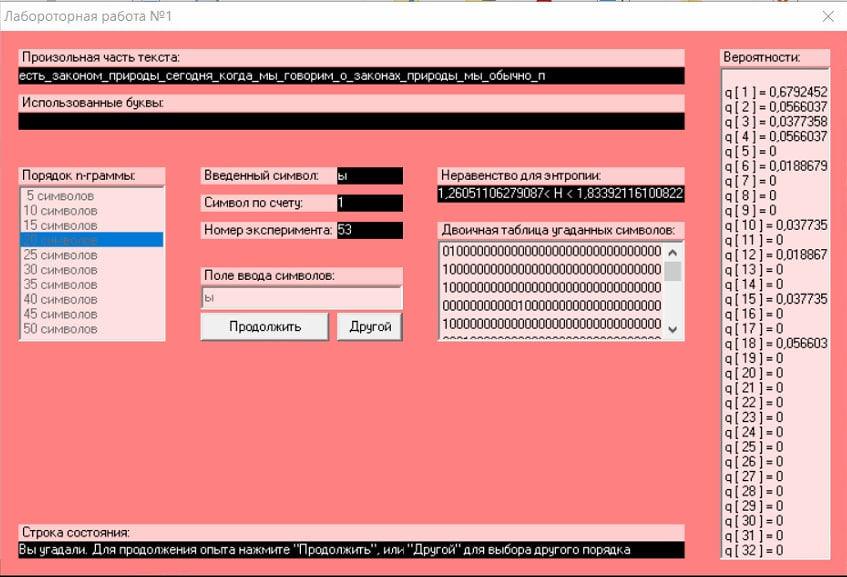
Для зручності перегляду були створені таблиці

**Робота з CoolPinkProgram**

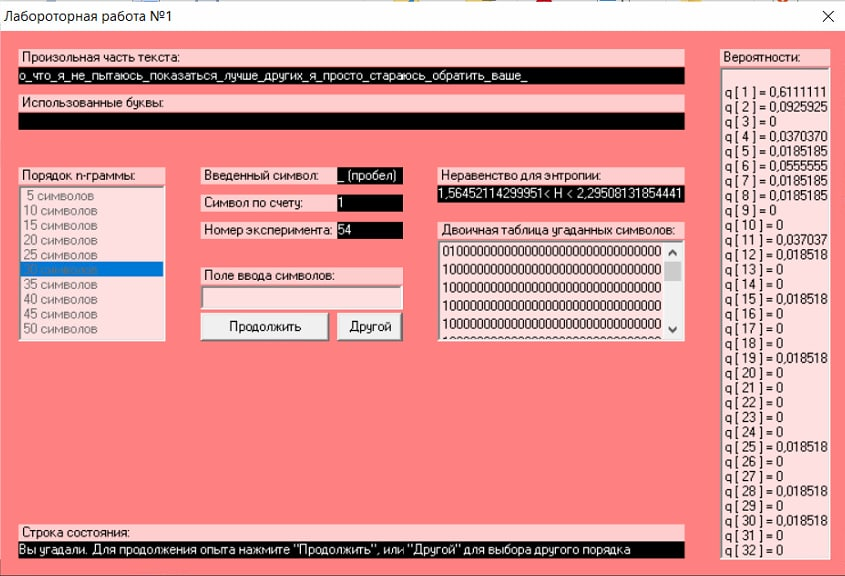
1,18<H(10) <1,75



1,26<H(20) <1,83



1,56<H(30) <2,29



**Оцінка надлишковості мови:**

|  |  |  |
| --- | --- | --- |
|  | R(Текст з пробілами) | R(Текст без пробілів) |
| Н1 | **0.1331179814452792** | **0.11466156950060935** |
| Н2 | **0.20433902779205404** | **0.1739815806779923** |
| Н2(з кроком 2) | **0.20441786826983577** | **0.1741072770900105** |

**Висновки:**

У ході виконання лабораторної роботи ми ознайомились з поняттям ентропії та навчились її визначати, освоїли вимірювання частоти повторювання символів в тексті. Ми дійшли такого висновку, що коли оброблений текст містить пробіл, ентропія є меншою. Надлишковість існує в межах від 0,11-0,2.