Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Виконала:
студентка
групи ФБ-13
Теплякова Анна

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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 1 *H* та 2 *H* за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 *H* та 2 *H* на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 *H* та 2 *H* 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.

Обраховані за допомогою написаного коду (lab1.py) значення ентропії:

Для тексту з пробілами:
H1 для тексту з пробілами: 4.367271375829809
R = 12.654572483403825 %
H2 для тексту з пробілами (біграми перетинаються): 7.915428210063267
R = 16.625777036701926 %
H2 для тексту з пробілами (біграми не перетинаються, з кроком 2): 7.914868194713824
R = 16.625777036701926 %

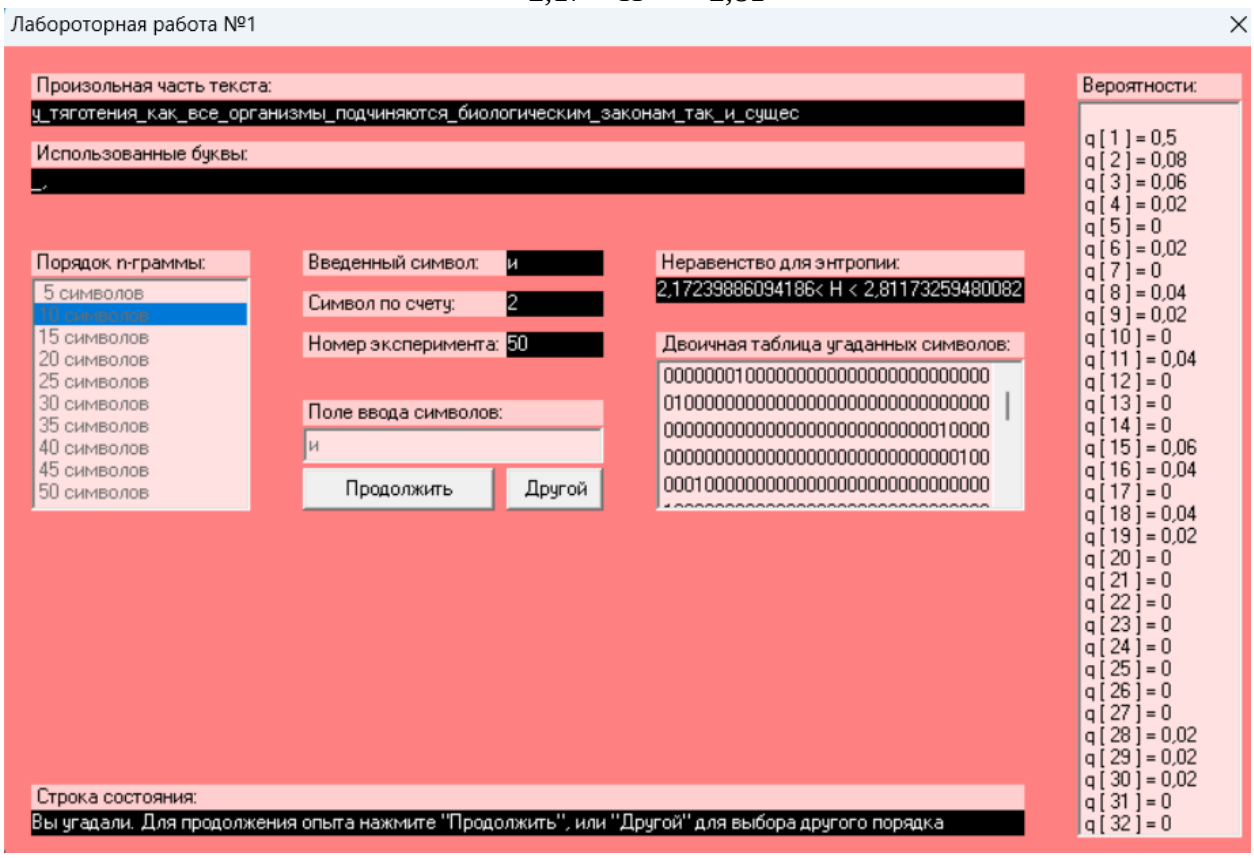
Для тексту з видаленими пробілами:
H1 без пробілів: 4.468635200737929
H2 без пробілів (біграми перетинаються): 8.304629606275359

Таблиці кількості та частоти літер та біграм в прикладеному текстовому файлі, бо щось вони занадто довгі для нормального виду протоколу

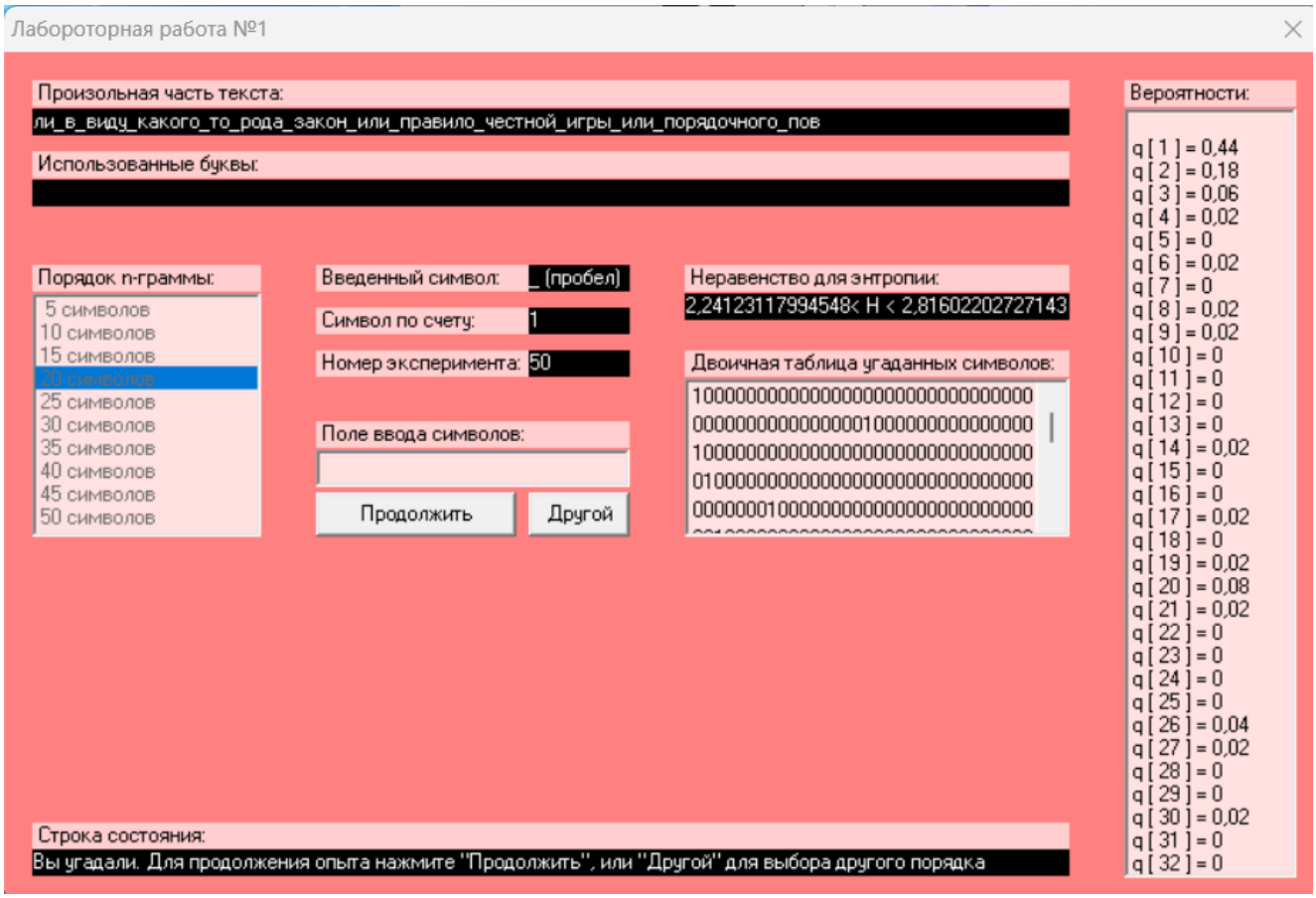
Перейдемо до роботи з CoolPinkProgram

Округлю значення ентропії

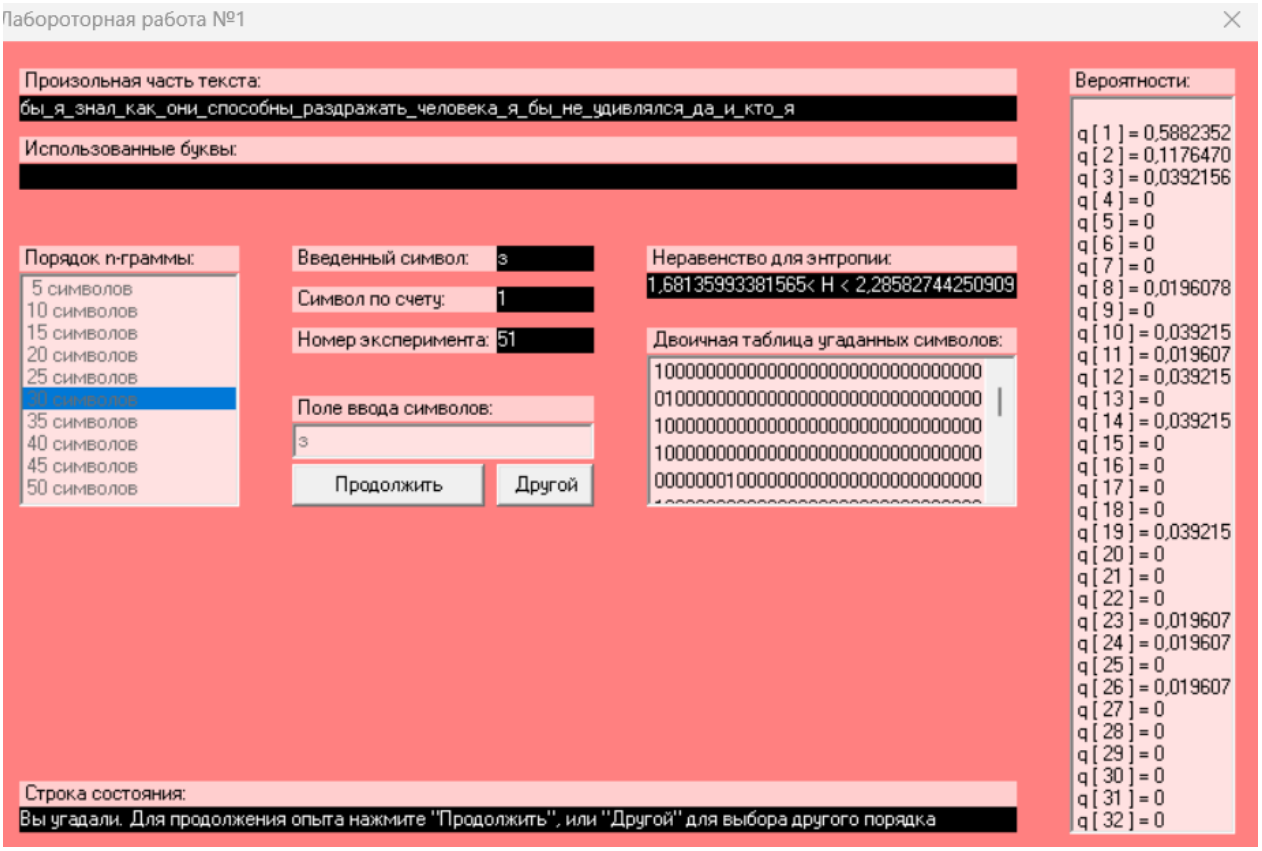
$2{,}17 < H^{(10)} < 2{,}81$



$2{,}241 < H^{(20)} < 2{,}816$



$$1{,}681 < H^{(30)} < 2{,}286$$



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

$$R = 1 - \frac{H_0}{H_\infty}$$

$H_0 = \log_2 32 = 5$

$H^{(10)}$

$$1 - \frac{2{,}17}{5} < R < 1 - \frac{2{,}81}{5}$$

$0{,}566 < R < 0{,}438$

$H^{(20)}$

$$1 - \frac{2{,}241}{5} < R < 1 - \frac{2{,}816}{5}$$

$0{,}5518 < R < 0{,}4368$

$H^{(30)}$

$$1 - \frac{1{,}681}{5} < R < 1 - \frac{2{,}286}{5}$$

$0{,}6638 < R < 0{,}5428$

Висновки:
Під час виконання роботи, я навчилася практичним шляхом підраховувати частоти букв та біграм на обраному тексті, обраховувати значення ентропії на символ джерела та надлишковість джерела відкритого тексту. За думкою Інтернету, надлишковість російської мови в цілому = 72,6%, а літературної мови = 76,2%, нехай середня надлишковість буде 74%. Моє найвище значення нижче на майже 10 відсотків. Це можна пояснити тим, що в усіх трьох експериментах я вгадувала літери з першого разу найбільше за інші літери (хоча я майже через раз просто починала клацати вздовж по клавіатурі). Бо чим непередбачуваніше повідомлення, тим більша його ентропія.