Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут ім. Ігоря Сікорського”

Фізико-технічний інститут

Лабораторна робота № 1

з предмету «Криптографія»

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:

Студент 3 курсу,

ФТІ, групи ФБ-05

Супрун Максим

Київ - 2021

**Мета:** 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.

**Постановка задачі:**

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

Для виконання даного завдання був використаний текст книги “віднесені вітром”, який був попередньо оброблений з метою виключення всіх символів, що не входять в початковий алфавіт. Текст називаэться sample\_text.txt .

Основною проблемою, яка зайняла найбільше часу, стали рішення при проєктуванні програми, що були покликані зменшити час виконання програми.

Повний код можна переглянути у відповідному [репозиторію на Гітхабі](https://github.com/minako-aino/crypto-FB-9/tree/cp1/cp1/dorosh_fb92_shatkovska_fb92_cp1).

Нижче наведено таблиці, які демонструють виконання коду.

**Результати:**

В даному прикладі я продемострував 10 найчастіших біграм, повну таблицю можна переглянути в репозиторії github

**З пробілами:**

“\_” - пробіл

**Частота літер у тексті:**

|  |  |
| --- | --- |
|  | 0,171524 |
| а | 0,071128 |
| б | 0,015203 |
| в | 0,034257 |
| г | 0,014619 |
| д | 0,025419 |
| е | 0,073094 |
| ж | 0,009436 |
| з | 0,013002 |
| и | 0,05478 |
| й | 0,008239 |
| к | 0,028774 |
| л | 0,044022 |
| м | 0,026241 |
| н | 0,054306 |
| о | 0,089762 |
| п | 0,020338 |
| р | 0,033058 |
| с | 0,043354 |
| т | 0,055138 |
| у | 0,023214 |
| ф | 0,000639 |
| х | 0,007016 |
| ц | 0,002094 |
| ч | 0,013093 |
| ш | 0,007375 |
| щ | 0,002349 |
| ъ | 0,000212 |
| ы | 0,015587 |
| ь | 0,016915 |
| э | 0,004001 |
| ю | 0,004642 |
| я | 0,017165 |

**Частота біграм, що перетинаються:**

|  |  |
| --- | --- |
| а | 0,021928 |
| о | 0,021434 |
| и | 0,020708 |
| е | 0,020441 |
| н | 0,016956 |
| с | 0,016469 |
| в | 0,015737 |
| п | 0,014443 |
| то | 0,012872 |
| о | 0,012091 |

В даному прикладі я продемострував 10 найчастіших біграм, повну таблицю можна переглянути в репозиторії github

**Частота біграм, що не перетинаються:**

|  |  |
| --- | --- |
| а | 0,021943 |
| о | 0,02151 |
| и | 0,020755 |
| е | 0,020361 |
| н | 0,016809 |
| с | 0,016536 |
| в | 0,015784 |
| п | 0,014402 |
| то | 0,012638 |
| о | 0,012035 |

**Без пробілів:**

**Частота літер у тексті:**

|  |  |
| --- | --- |
| а | 0,085855 |
| б | 0,01835 |
| в | 0,04135 |
| г | 0,017646 |
| д | 0,030682 |
| е | 0,088227 |
| ж | 0,01139 |
| з | 0,015694 |
| и | 0,066122 |
| й | 0,009945 |
| к | 0,034732 |
| л | 0,053137 |
| м | 0,031673 |
| н | 0,065549 |
| о | 0,108346 |
| п | 0,024549 |
| р | 0,039903 |
| с | 0,05233 |
| т | 0,066553 |
| у | 0,028021 |
| ф | 0,000772 |
| х | 0,008469 |
| ц | 0,002528 |
| ч | 0,015804 |
| ш | 0,008902 |
| щ | 0,002835 |
| ъ | 0,000256 |
| ы | 0,018815 |
| ь | 0,020417 |
| э | 0,004829 |
| ю | 0,005603 |
| я | 0,020719 |

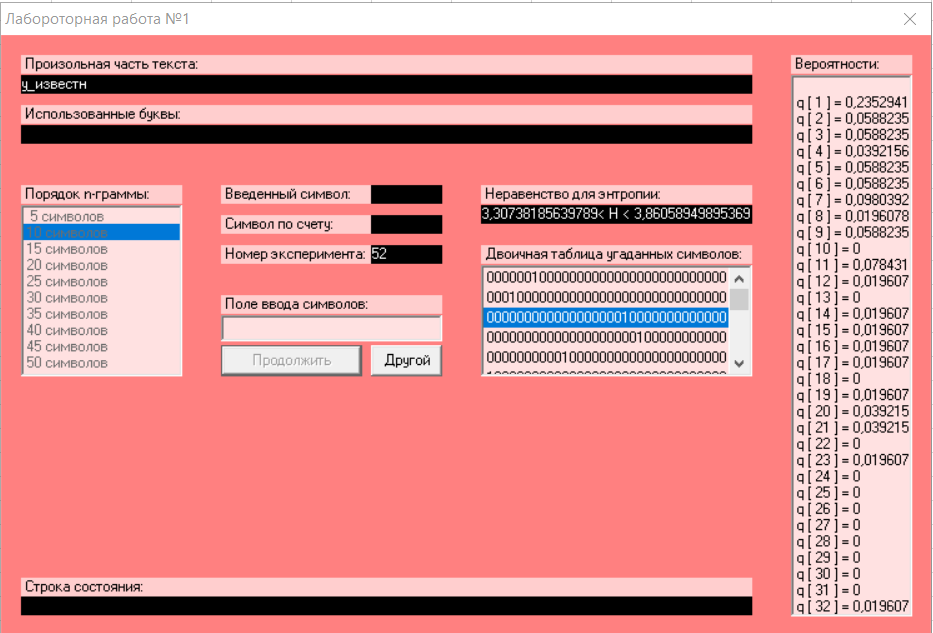
**Частота біграм, що перетинаються:**

|  |  |
| --- | --- |
| то | 0,01621 |
| на | 0,013311 |
| не | 0,013144 |
| ла | 0,012442 |
| он | 0,012202 |
| ст | 0,011456 |
| по | 0,010611 |
| но | 0,010399 |
| ка | 0,010337 |
| ет | 0,009881 |

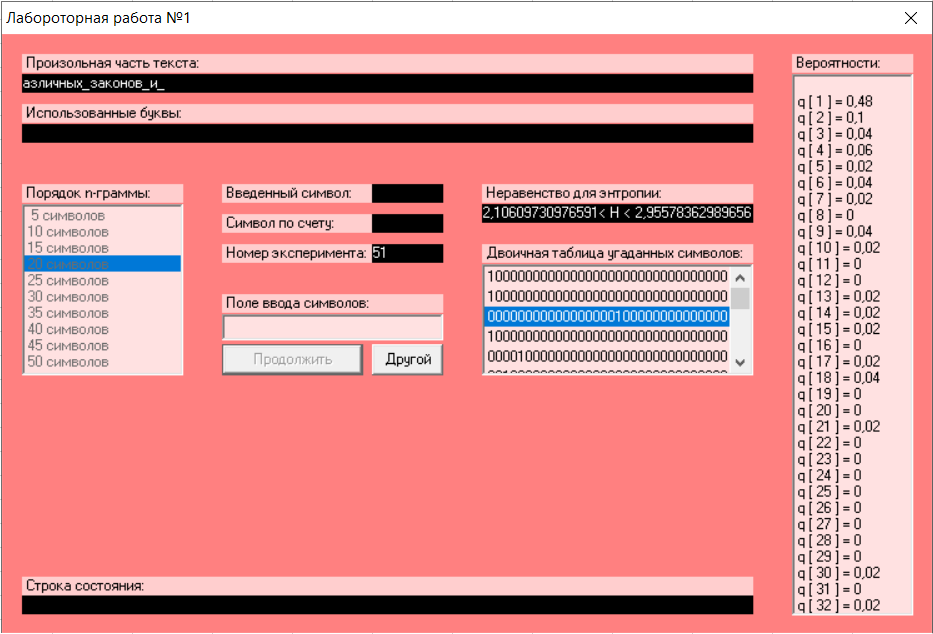
**Частота біграм, що не перетинаються:**

|  |  |
| --- | --- |
| то | 0,015771 |
| на | 0,013254 |
| не | 0,013071 |
| ла | 0,01257 |
| он | 0,011889 |
| ст | 0,011557 |
| по | 0,010854 |
| но | 0,010587 |
| ка | 0,010203 |
| ет | 0,009936 |

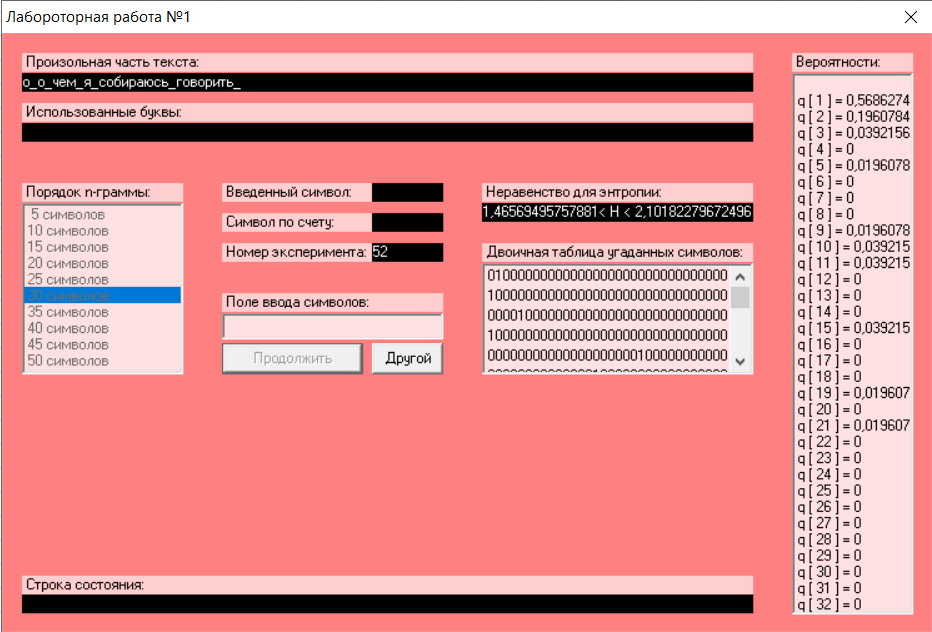
**Значення ентропії:**



0,349895624 < R < 0,241156227



0,586022075 < R < 0,419006345



0,711900607 < R < 0,58686228

**Вмсновок**

При виконанні роботи навчався рахувати частоту символів та біграм у тексті, оцінювати надлишковість російської мови, та рахувати ентропію. Зі збільшенням значення довжини тексту було помітно, що умовна ентропія зростає та ми маємо більше і більше інформації про наступний символ.