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:**

|  |  |
| --- | --- |
|  | 0,162539 |
| а | 0,067403 |
| б | 0,014406 |
| в | 0,032463 |
| г | 0,013853 |
| д | 0,024088 |
| е | 0,069265 |
| ж | 0,008942 |
| з | 0,012321 |
| и | 0,051911 |
| й | 0,007808 |
| к | 0,027267 |
| л | 0,041716 |
| м | 0,024866 |
| н | 0,051461 |
| о | 0,08506 |
| п | 0,019273 |
| р | 0,031327 |
| с | 0,041083 |
| т | 0,05225 |
| у | 0,021998 |
| ф | 0,000606 |
| х | 0,006649 |
| ц | 0,001985 |
| ч | 0,012407 |
| ш | 0,006989 |
| щ | 0,002226 |
| ъ | 0,000201 |
| ы | 0,014771 |
| ь | 0,016029 |
| э | 0,003791 |
| ю | 0,004399 |
| я | 0,016266 |

**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 | 0,010972 |
| о | 0,010755 |
| и | 0,010377 |
| е | 0,01018 |
| н | 0,008404 |
| с | 0,008268 |
| в | 0,007892 |
| п | 0,007201 |
| то | 0,006319 |
| о | 0,006018 |

**Без пробілів:**

**Частота літер у тексті:**

|  |  |
| --- | --- |
| а | 0,067403 |
| б | 0,014406 |
| в | 0,032463 |
| г | 0,013853 |
| д | 0,024088 |
| е | 0,069265 |
| ж | 0,008942 |
| з | 0,012321 |
| и | 0,051911 |
| й | 0,007808 |
| к | 0,027267 |
| л | 0,041716 |
| м | 0,024866 |
| н | 0,051461 |
| о | 0,08506 |
| п | 0,019273 |
| р | 0,031327 |
| с | 0,041083 |
| т | 0,05225 |
| у | 0,021998 |
| ф | 0,000606 |
| х | 0,006649 |
| ц | 0,001985 |
| ч | 0,012407 |
| ш | 0,006989 |
| щ | 0,002226 |
| ъ | 0,000201 |
| ы | 0,014771 |
| ь | 0,016029 |
| э | 0,003791 |
| ю | 0,004399 |
| я | 0,016266 |

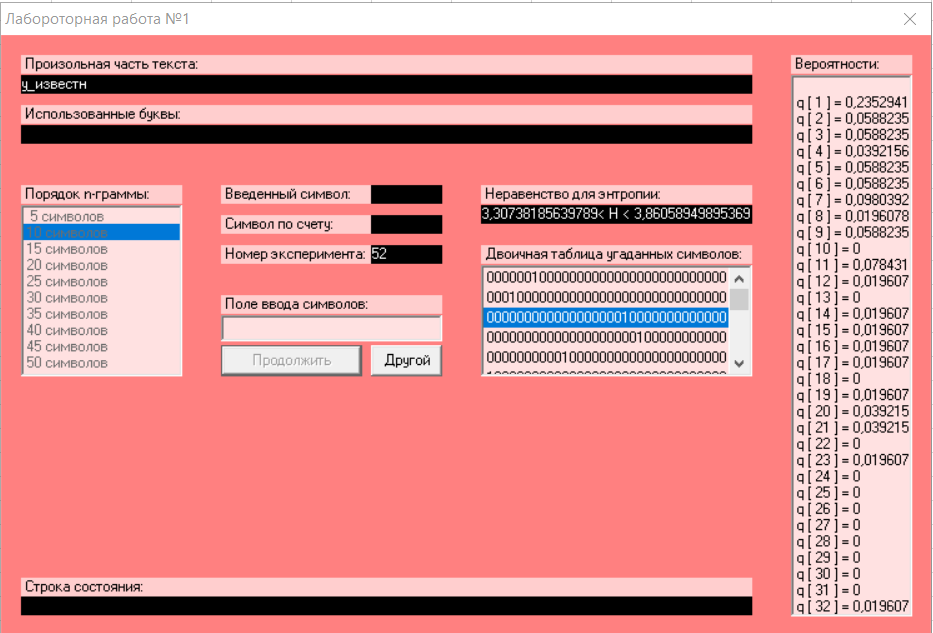
**Частота біграм, що перетинаються:**

|  |  |
| --- | --- |
| 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 | 0,010336 |
| ет | 0,009881 |

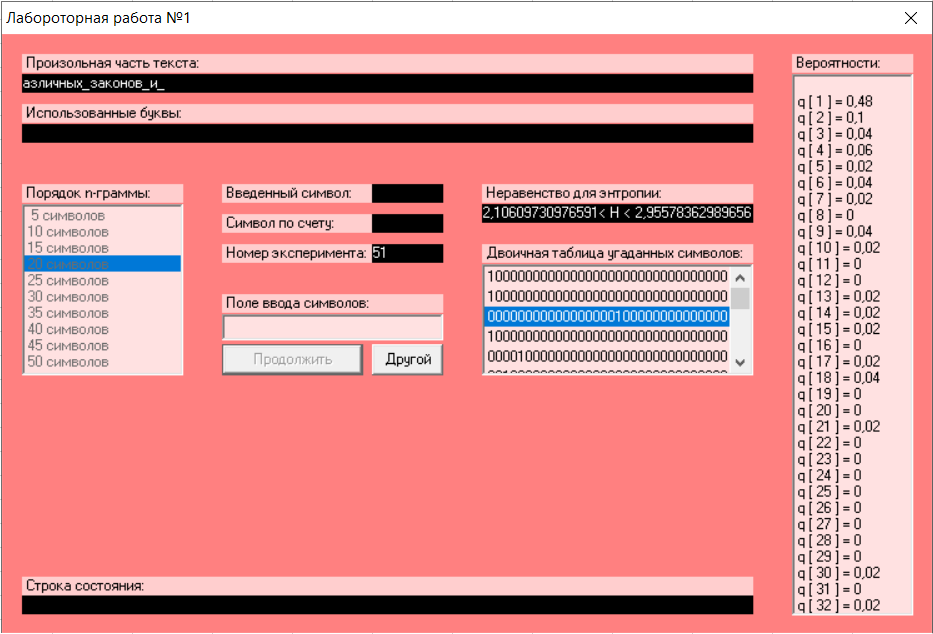
**Частота біграм, що не перетинаються:**

|  |  |
| --- | --- |
| то | 0,007885 |
| на | 0,006627 |
| не | 0,006536 |
| ла | 0,006285 |
| он | 0,005945 |
| ст | 0,005778 |
| по | 0,005427 |
| но | 0,005294 |
| ка | 0,005101 |
| ет | 0,004968 |

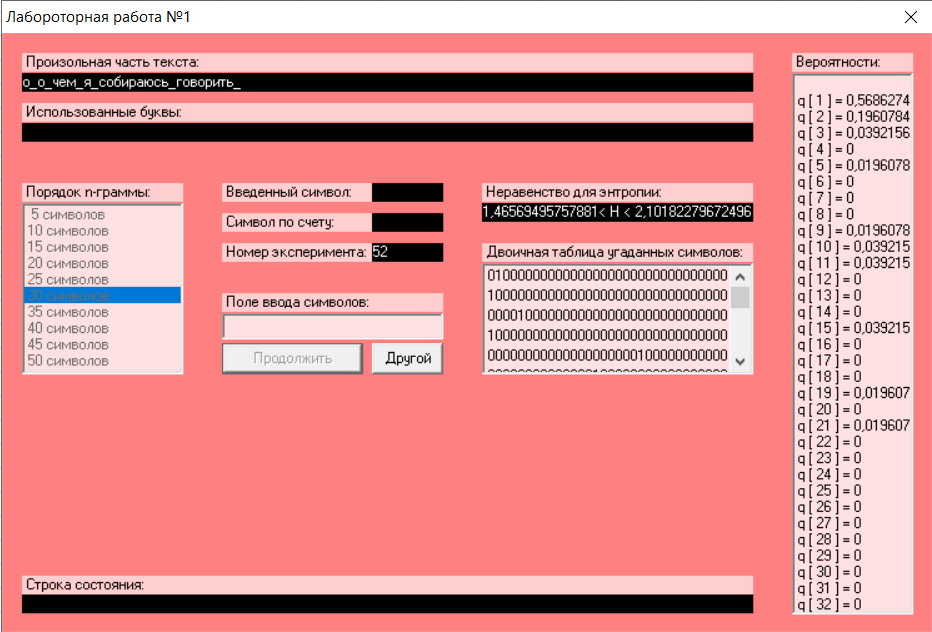
**Значення ентропії:**



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.