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 текст

Виконав

Студенти гр.Фб-22 Пунько Артем

гр.Фб-22 Руденко Поліна

Київ – 2024

**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.

**Постановка задачі:**

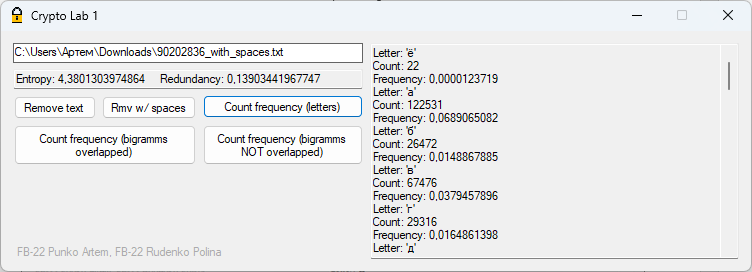
1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку 𝐻1 та 𝐻2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 𝐻1 та 𝐻2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 𝐻1 та 𝐻2 на тому ж тексті, в якому вилучено всі пробіли.
2. За допомогою програми CoolPinkProgram оцінити значення 𝐻 (10) , 𝐻 (20) , 𝐻 (30) .
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

**Хід роботи:**

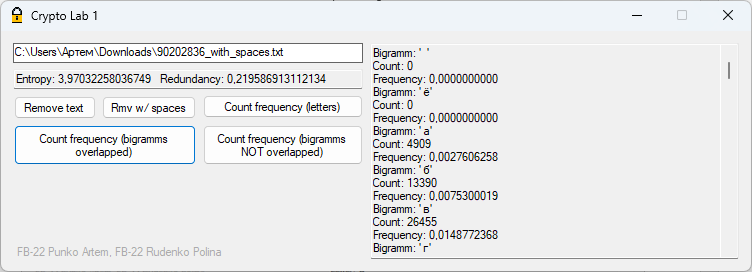
Щоб виконати лабороторну роботу було розроблено програму, для аналізу тексту та обчислення різних статистичних показників. Загальні етапи для роботи коду були: Зчитавання тексту з файлу, попердня обробку текста, створення нової версії тексту без пробілів, підрахунок частос символів та біграм, обчислення ентропії H1 та H2, обчислення падлишковості, збереження результатів у файлі.

**Аналіз тексту з пробілами**

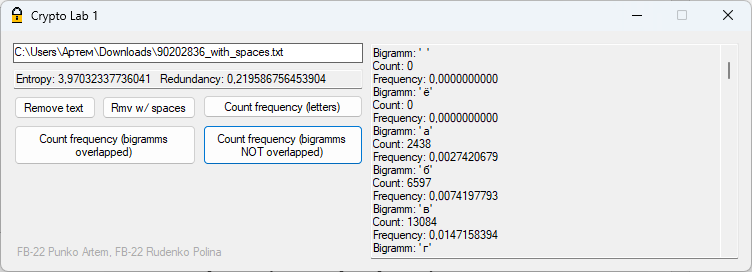
**Літери**

****

**Біграми (з перекриттям)**

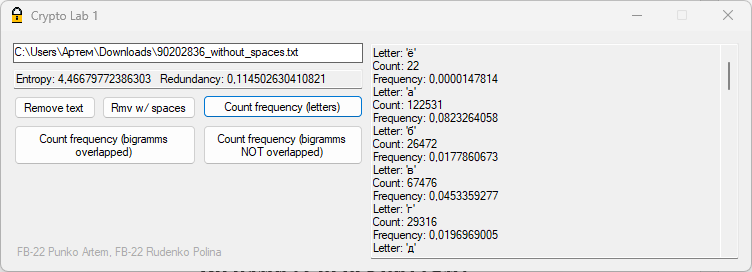
****

**Біграми (без перекритя)**

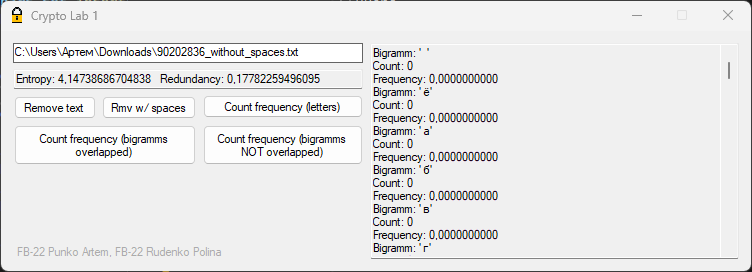
****

**Аналіз тексту БЕЗ пробілів**

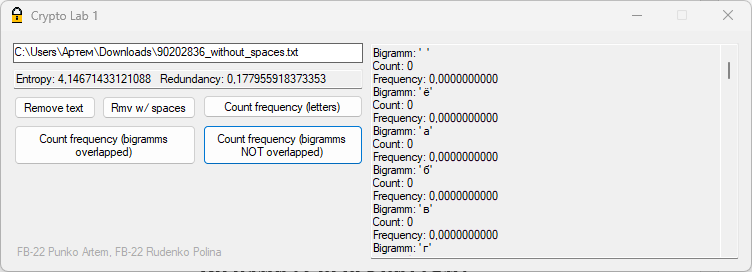
**Літери**

****

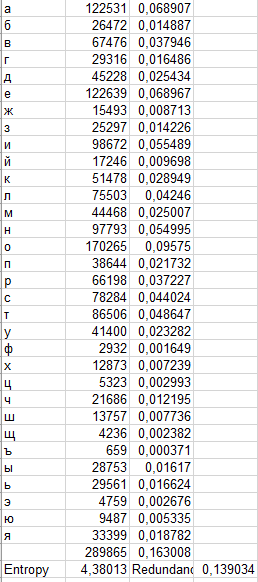
**Біграми (з перекриттям)**

****

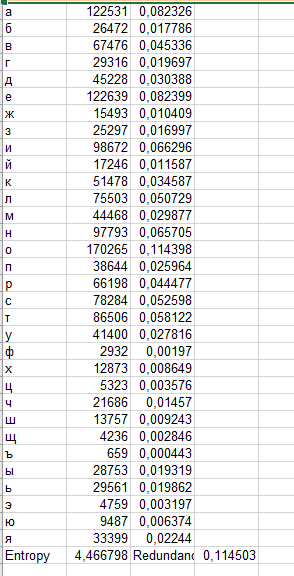
**Біграми (без перекритя)**

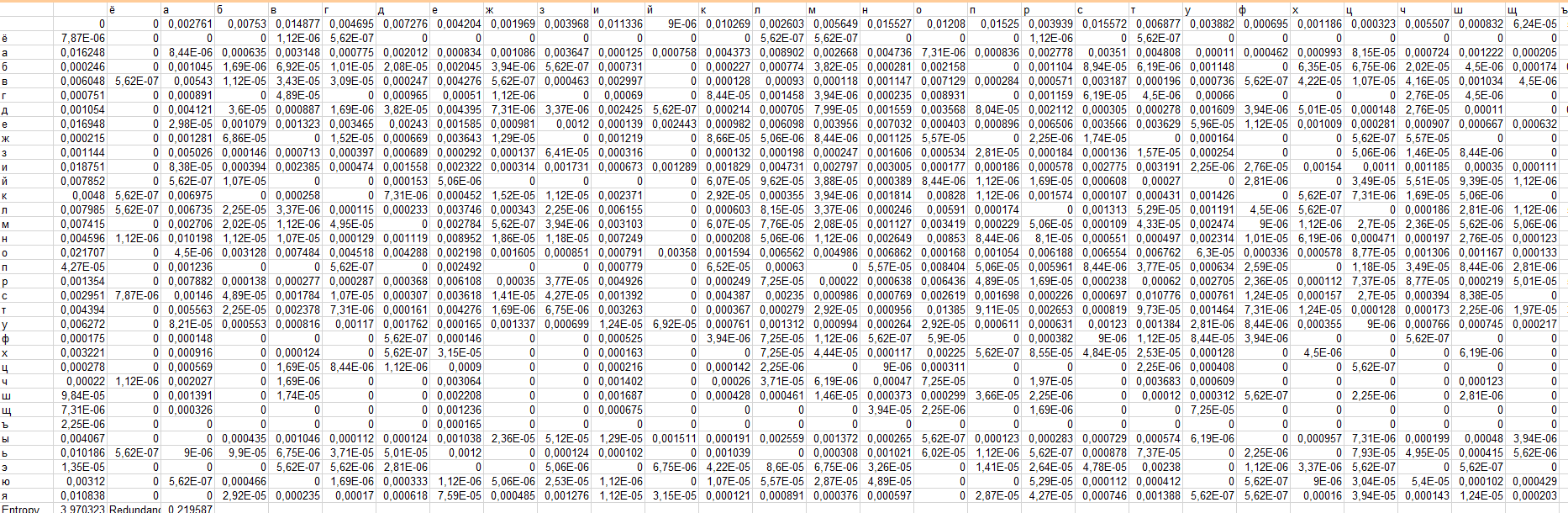
****

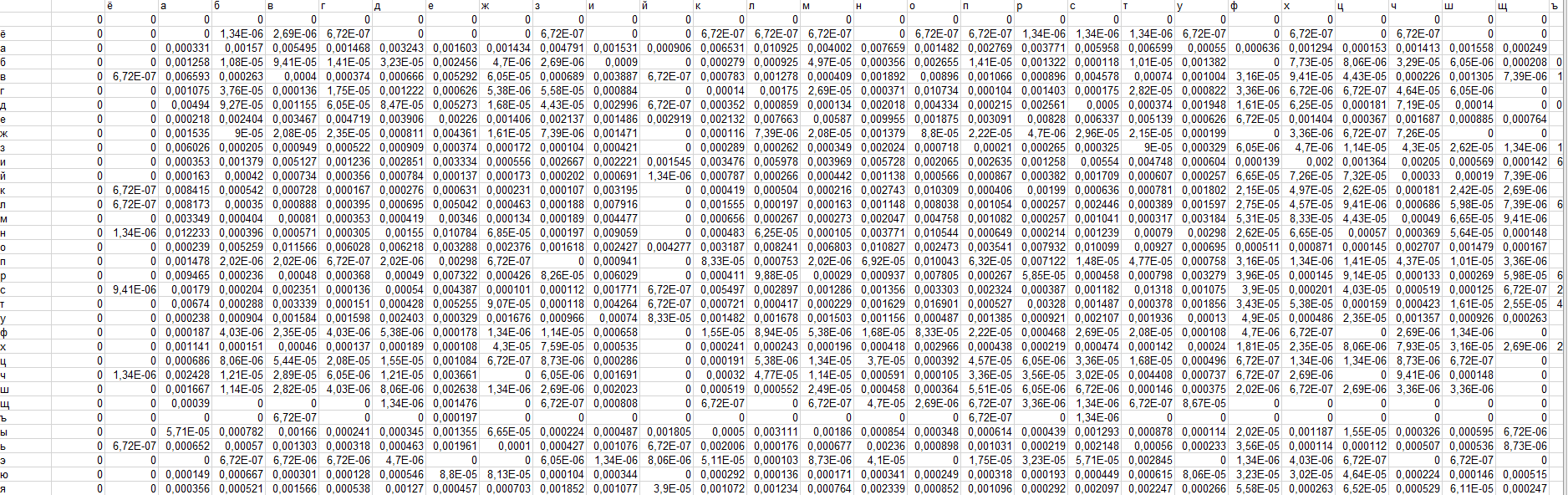
**Таблиця з частотами літер з пробілом**

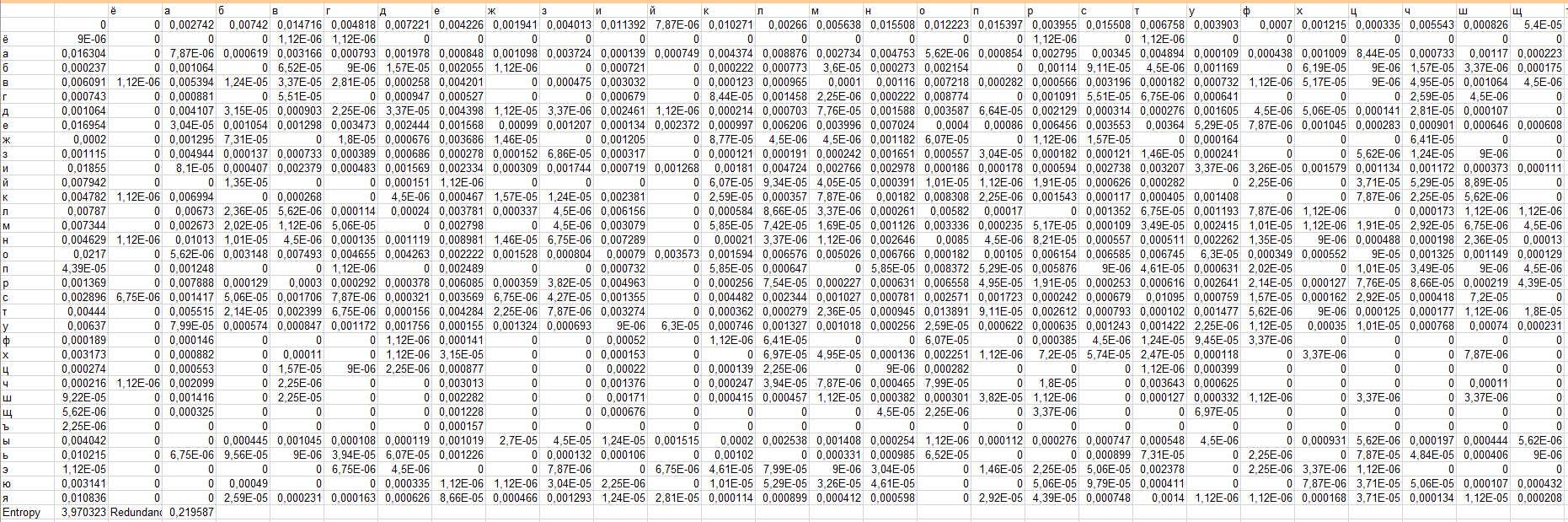
****

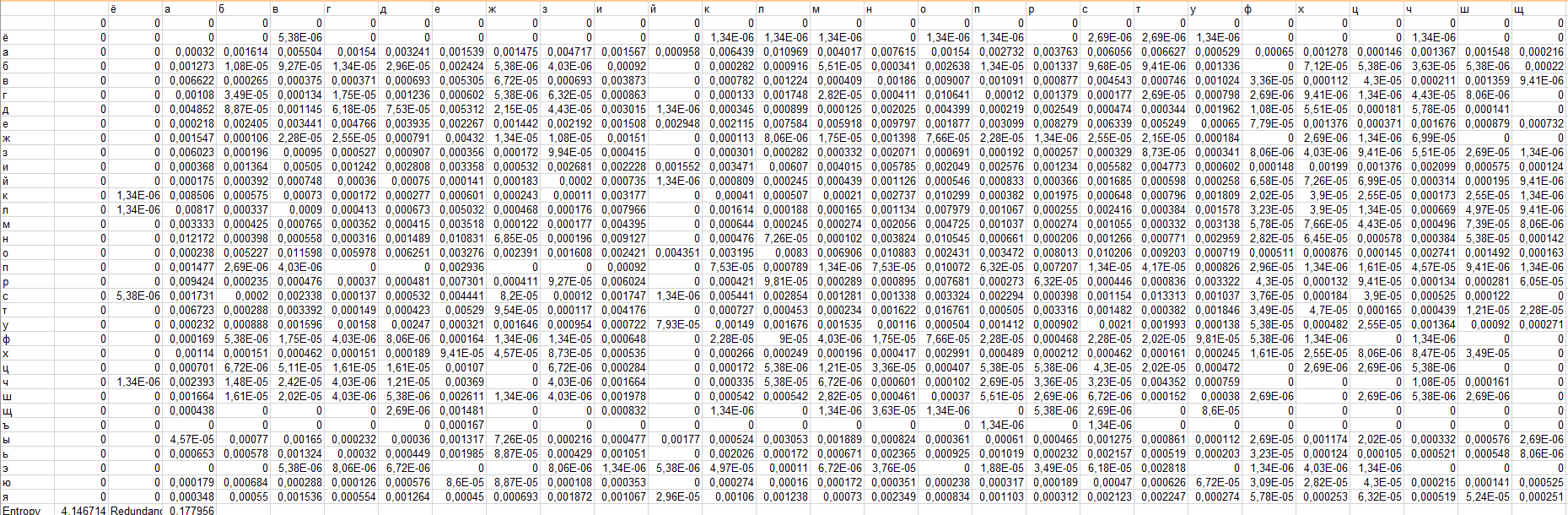
**Таблиця з частотами літер без пробілу**

****

**Таблиця з частотами біграм з перекриттям з пробілами**

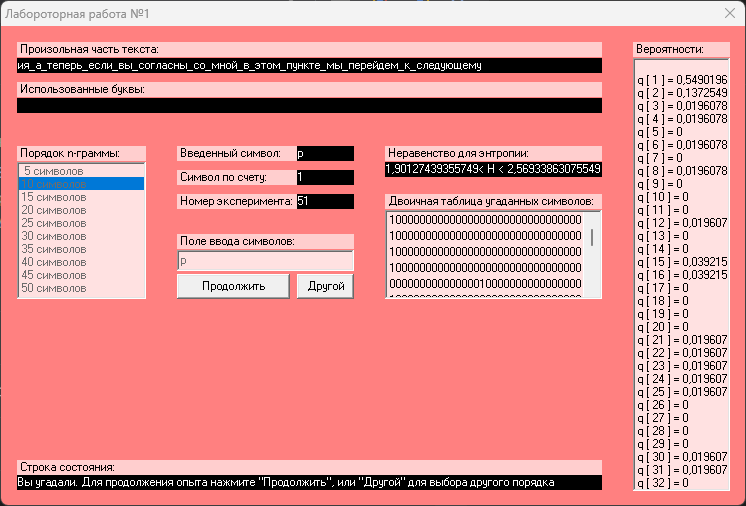
**Таблиця з частотами біграм з перекриттям без пробілів**

** Таблиця з частотами біграм без перекриття з пробілами**

**Таблиця з частотами біграм без перекриття без пробілів**

|  |  |  |
| --- | --- | --- |
| **Тип вибірки** | **H** | **R** |
| H1 з пробілами | 4,3801303974864 | 0,13903441967747 |
| H1 без пробілів | 4,46679772386303 | 0,114502630410821 |
| H2 з пробілами і перекриттям | 3,97032258036749 | 0,219586913112134 |
| H2 без пробілів і з перекриттям | 4,14738686704838 | 0,17782259496095 |
| H2 з пробілами і без перекриттям | 3,97032337736041 | 0,219586756453904 |
| H2 без пробілів і без перекриттям | 4,14671433121088 | 0,177955918373353 |

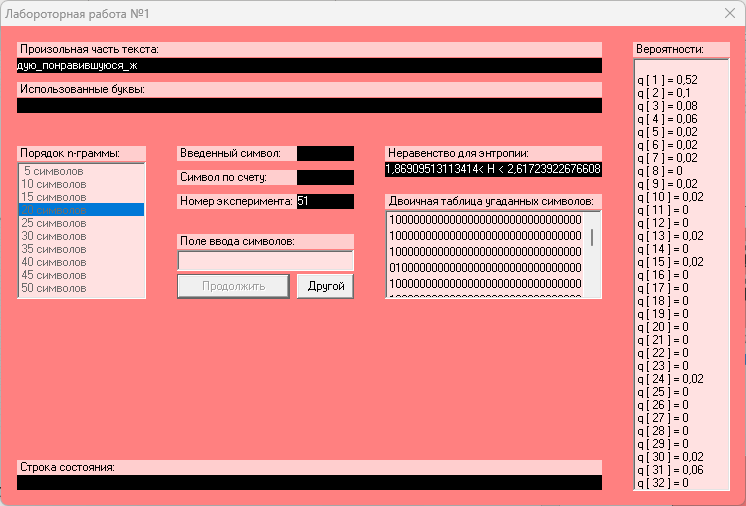
**Н(10)**

****

1.90127439355749 < H(10) < 2.56933863075549

0.48613227384 < R(10) < 0.61974512128

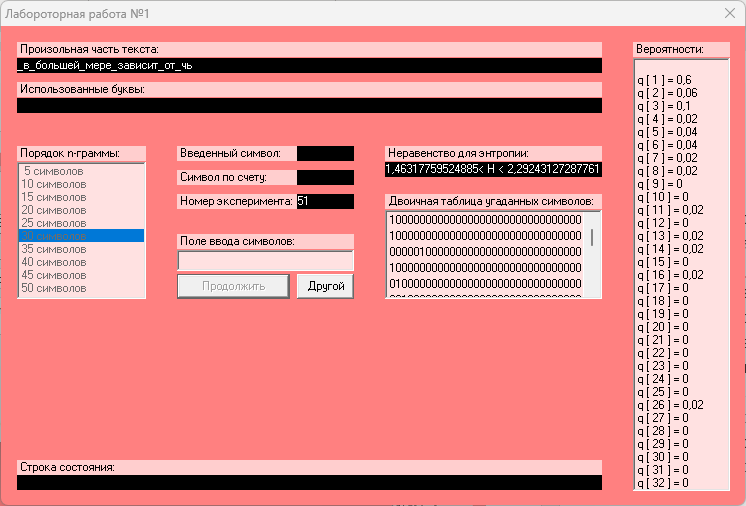
**H(20)**

****

1.86909513113414 < H(20) < 2.61723922676608

0.47655215464 < R(20) < 0.62618097377

**H(30)**

****

1.46317759524885 < H(20) < 2.29243127287761

0.54151374542 < R(20) < 0.70736448095

**Висновки**

Під час та після виконання роботи, ми зрозуміли що являє собою поняття «ентропія» в інформаційному сенсі (чим більше ентропія Hn – тим складніше вгадати блок інформації, яку утримує в собі n-грама) і навчилися її обраховувати. Також було засвоєно поняття надлишковості, що означає собою передбачуваність мови, тобто показник R являє собою наскільки відсотків де-яка моя є передбачуваною, тобто людина чи комп’ютер з шансом R може вгадати наступну літеру даної мови.

Також, за підсумками роботи можна заявити, що відсутність пробілу хоч і зменшує алфавіт мови, але піднімає ентропію, а значить і знижує надлишковість. Також, в моєму випадку, можна побачити дуже високу наближеність ентропії і надлишковості при відсутності і наявності перекриття в біграмах, що може свідчити про стабільну структуру тексту, тобто текст має не так багато унікальних біграм.