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:  
ФБ-21 Худоба Арсен,

ФБ-21 Шабанов Кирило

**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**

1. Написати програми для підрахунку a) частот букв і b) частот біграм в тексті, а також підрахунку c) 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 томуж тексті, в якому вилучено всі пробіли.

А) частоти букв

1. Частота літер

Частота окремої літери обчислюється як відношення кількості появ цієї літери до загальної кількості літер у тексті.

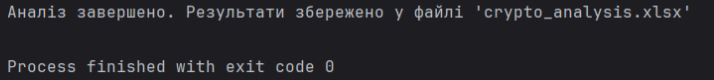
Нехай:

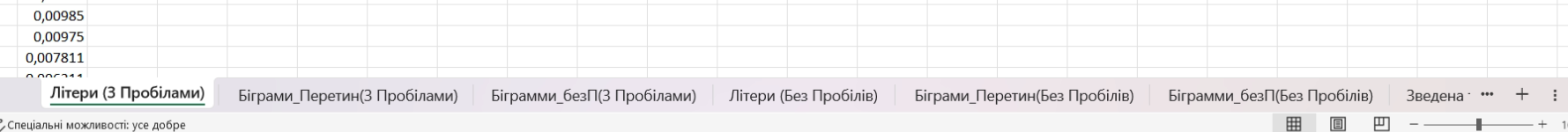
n — загальна кількість літер у тексті;

𝑓(𝑙) — кількість появ літери 𝑙 у тексті.

Тоді частота літери 𝑙 визначається як:

𝑃(𝑙) =

​ 



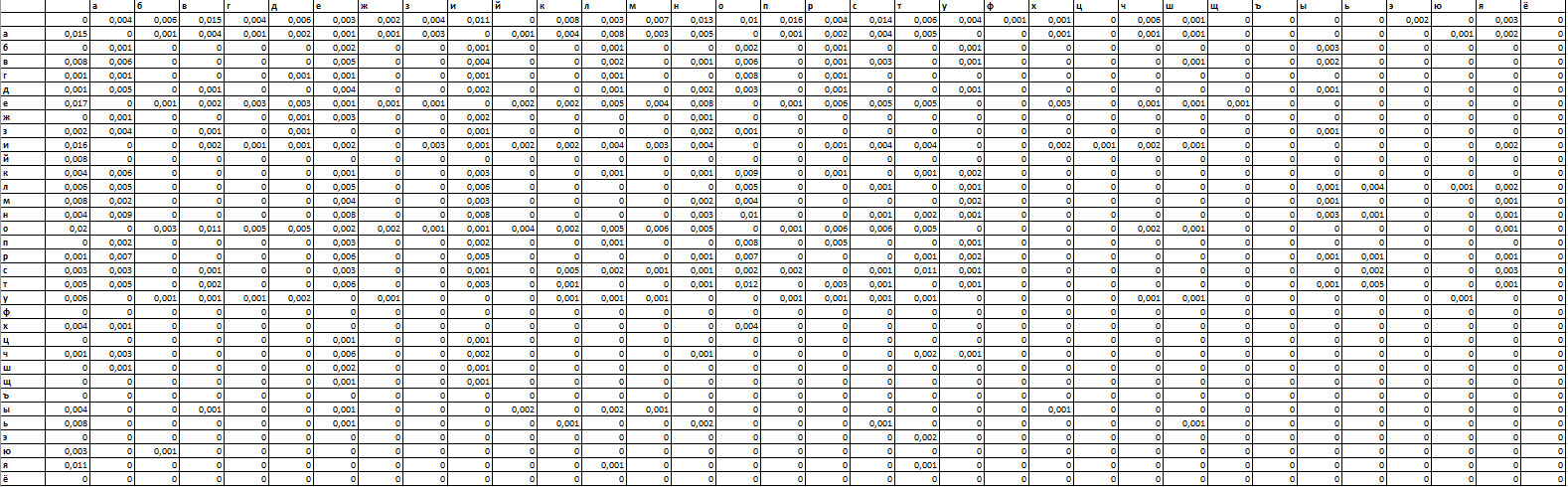
Частота літер з пробілом

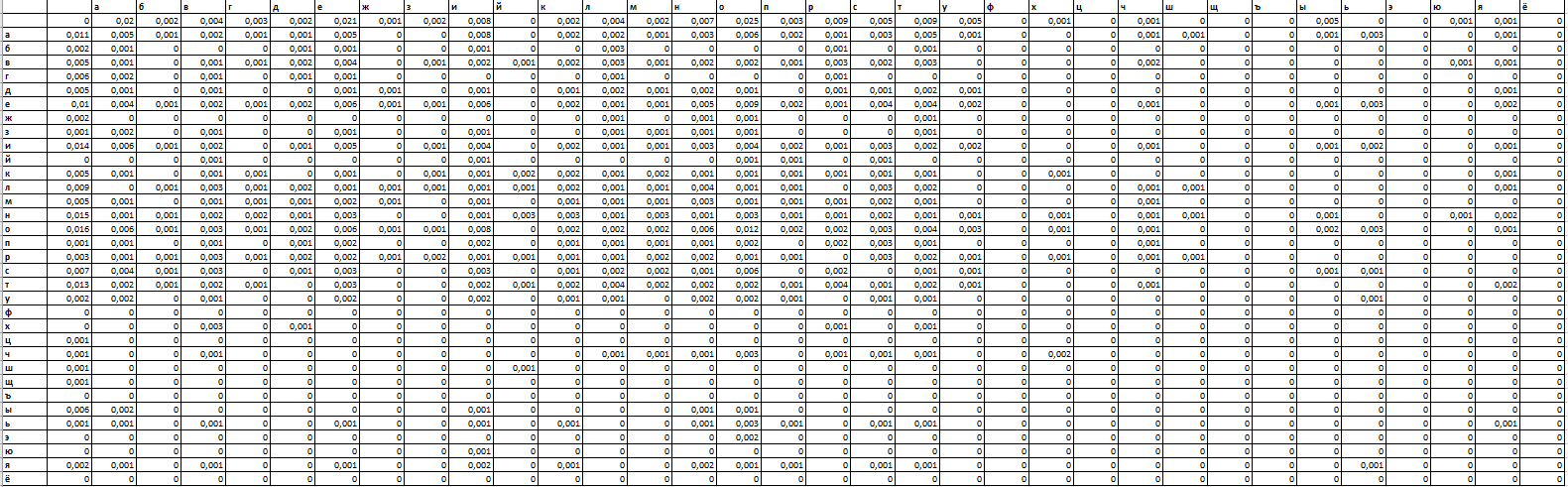
|  |  |
| --- | --- |
| **Літеру** | **Частота** |
|  | 0,1714408204 |
| о | 0,0926606594 |
| е | 0,0723308935 |
| а | 0,0667676716 |
| и | 0,0588510235 |
| н | 0,0537615605 |
| т | 0,0488305470 |
| с | 0,0451899169 |
| в | 0,0418063036 |
| л | 0,0384311693 |
| р | 0,0357942297 |
| к | 0,0288595448 |
| м | 0,0267641953 |
| д | 0,0233954201 |
| п | 0,0228368722 |
| у | 0,0211379115 |
| я | 0,0165227372 |
| ч | 0,0153807347 |
| г | 0,0153489388 |
| ь | 0,0147951602 |
| ы | 0,0147919807 |
| б | 0,0133595755 |
| з | 0,0131799287 |
| й | 0,0098503685 |
| х | 0,0097502114 |
| ж | 0,0078106618 |
| ш | 0,0063109554 |
| ю | 0,0048785503 |
| ц | 0,0027922096 |
| щ | 0,0025500308 |
| э | 0,0023380582 |
| ф | 0,0012956828 |
| ъ | 0,0001738176 |
| ё | 0,0000116585 |

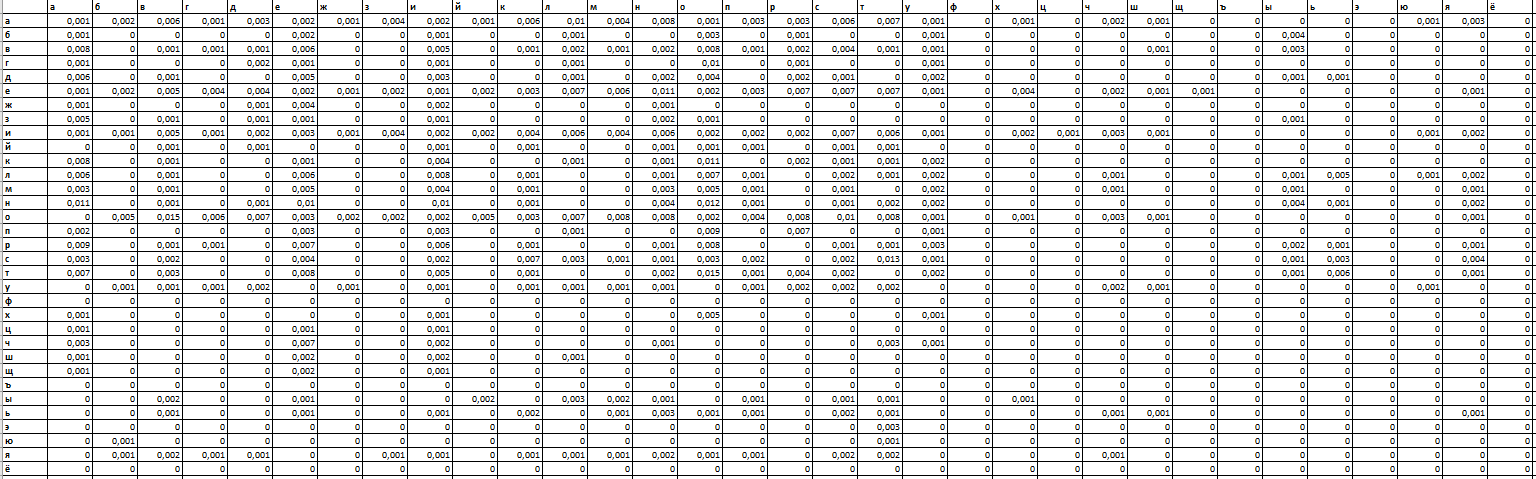
Частота літер без пробілу

|  |  |
| --- | --- |
| **Літеру** | **Частота** |
| о | 0,1118334835 |
| е | 0,0872971965 |
| а | 0,0805828639 |
| и | 0,0710281474 |
| н | 0,0648856012 |
| т | 0,0589342900 |
| с | 0,0545403611 |
| в | 0,0504566296 |
| л | 0,0463831314 |
| р | 0,0432005710 |
| к | 0,0348310000 |
| м | 0,0323020925 |
| д | 0,0282362693 |
| п | 0,0275621498 |
| у | 0,0255116497 |
| я | 0,0199415294 |
| ч | 0,0185632300 |
| г | 0,0185248551 |
| ь | 0,0178564919 |
| ы | 0,0178526544 |
| б | 0,0161238640 |
| з | 0,0159070457 |
| й | 0,0118885515 |
| х | 0,0117676705 |
| ж | 0,0094268002 |
| ш | 0,0076167829 |
| ю | 0,0058879926 |
| ц | 0,0033699579 |
| щ | 0,0030776689 |
| э | 0,0028218361 |
| ф | 0,0015637782 |
| ъ | 0,0002097829 |
| ё | 0,0000140708 |

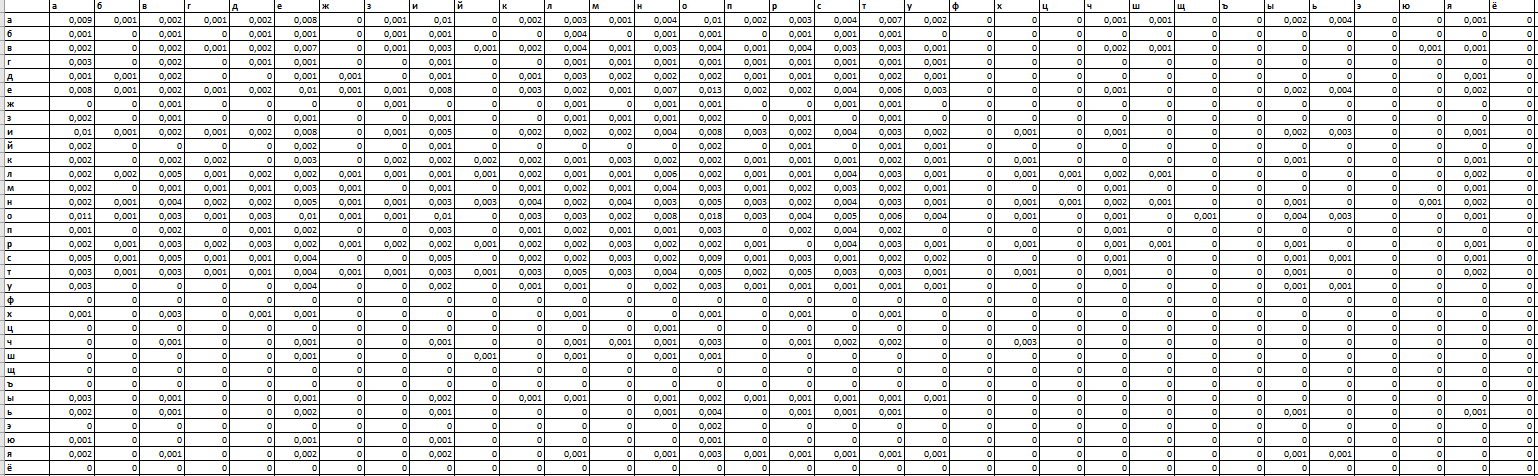
Біграми які перетинаються з пробілами

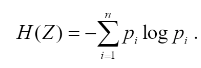
Біграми які не перетинаються з пробілами

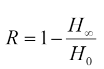


Біграми які перетинаються без пробілів

Біграми які не перетинаються з без пробілів



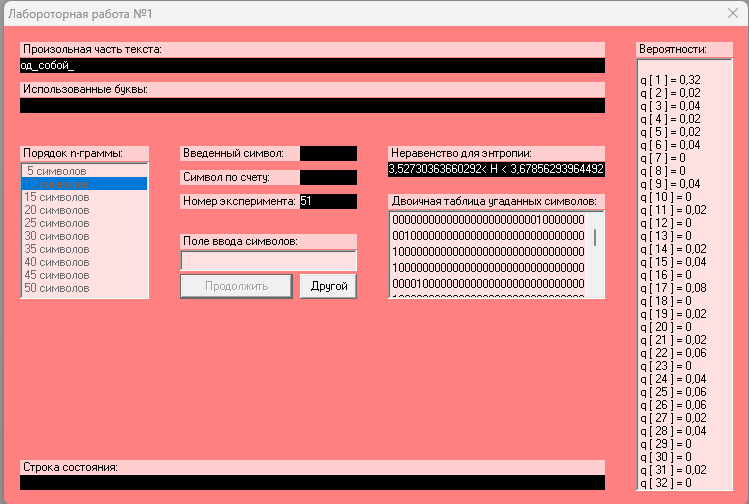


де 𝑛 – кількість символів у тексті; – частота i-того символу. 

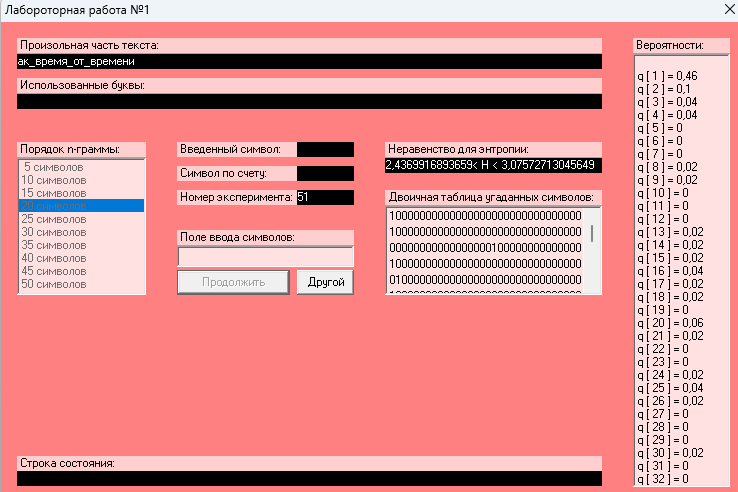
|  |  |  |
| --- | --- | --- |
| **Метрика** | **З Пробілами** | **Без Пробілів** |
| H1 | 4,34946 | 4,45166 |
| R1 | 0,14506 | 0,11750 |
| H2\_Перетин | 3,98588 | 4,14666 |
| R2\_Перетин | 0,29552 | 0,28897 |
| H2\_Без\_Перетин | 4,23692 | 4,35678 |
| R2\_Без\_Перетин | 0,28874 | 0,28151 |

2. За допомогою програми CoolPinkProgram оцінити значення

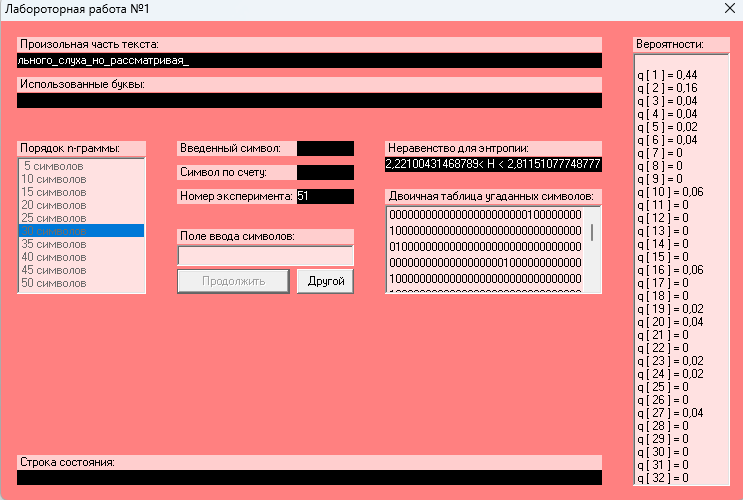
H (10)



H(20)



H(30)



|  |  |  |
| --- | --- | --- |
| Порядок n-грами | Ентропія | Надлишковість |
| 10 символів | 3.527< H < 3.678 | 0.27< R < 0.30 |
| 20 символів | 2.369< H < 3.075 | 0.39< R < 0.53 |
| 30 символів | 2.221< H < 2.811 | 0.44< R < 0.56 |

**Висновок**:

Дослідили та зрозуміли концепцію ентропії та надлишковості в контексті текстових даних. Реалізували підрахунок частот букв та біграм, створили матрицю біграм. Використали CoolPinkProgram, щоб експерементально виявити залежність ентропії та надлишковості від кількості букв.