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-31 Аль-Фітурі Асія  
ФБ-31 Гриб Вероніка

**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 тому ж тексті, в якому вилучено всі пробіли.

А) частота букв

1. Частота літер

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

𝑃(𝑙) =

Дані збереглись а табличці



Частота літери з пробілами

Зображення, що містить текст, знімок екрана, число, Шрифт

Вміст на основі ШІ може бути неправильним.

Частота літерів без пробілів

Зображення, що містить текст, знімок екрана, число, Шрифт

Вміст на основі ШІ може бути неправильним.

Біграми які перетинаются з пробілом

Зображення, що містить текст, чорно-білий, Шрифт, чорний

Вміст на основі ШІ може бути неправильним.

Біграми з пробілом без перетину

Зображення, що містить текст, знімок екрана, документ, Шрифт

Вміст на основі ШІ може бути неправильним.

Без пробілу з перетином

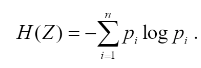
Зображення, що містить текст, знімок екрана, документ, Шрифт

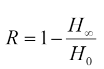
Вміст на основі ШІ може бути неправильним.

Без пробілу без перетину

Зображення, що містить текст, Шрифт, знімок екрана, документ

Вміст на основі ШІ може бути неправильним.

Обчислюємо ентропію символів за формулою: 

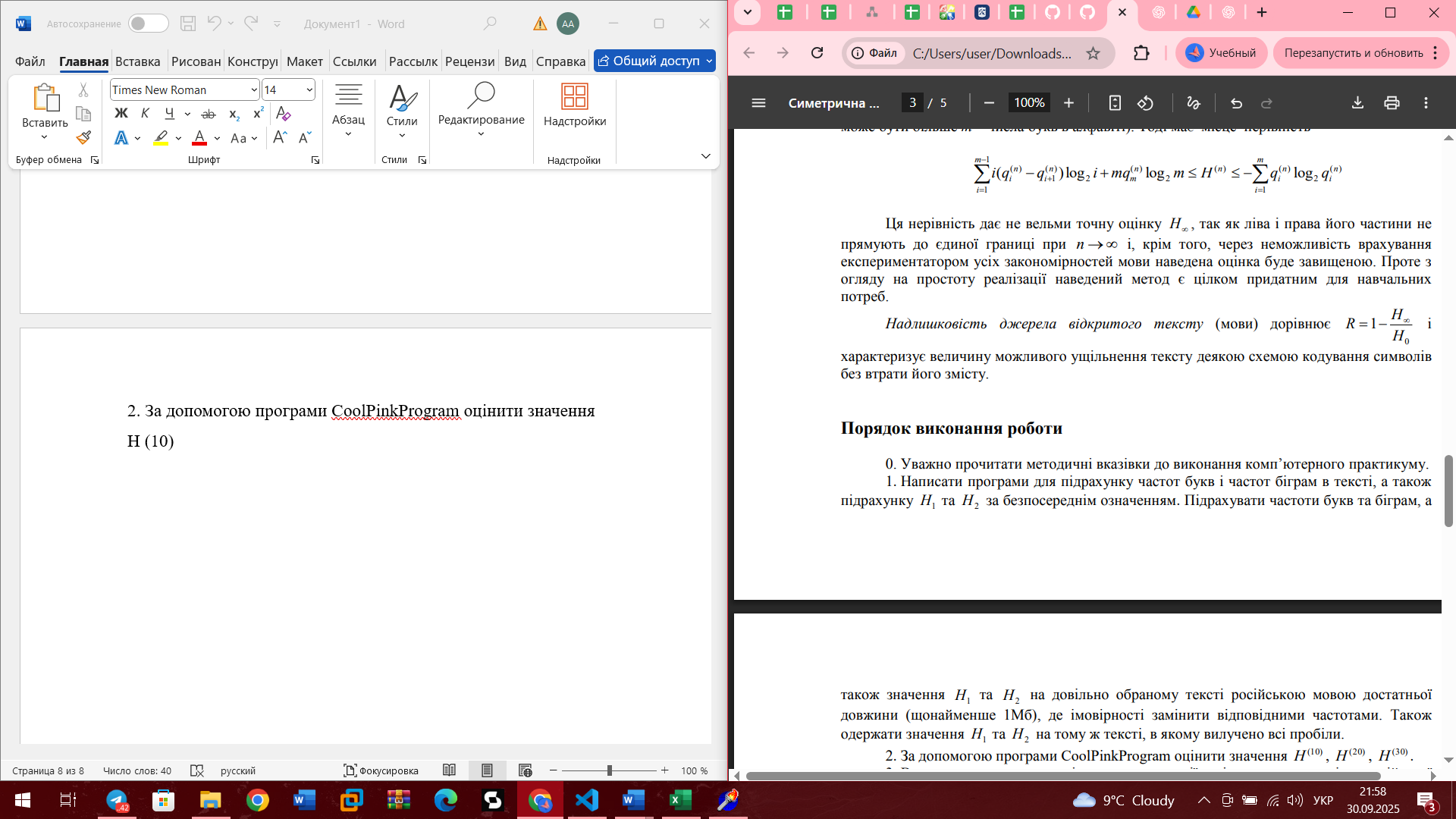
Надлишковість: 

Зображення, що містить текст, знімок екрана, Шрифт, число

Вміст на основі ШІ може бути неправильним.

2. За допомогою програми CoolPinkProgram оцінити значення

H (10), H (20) , H (30)



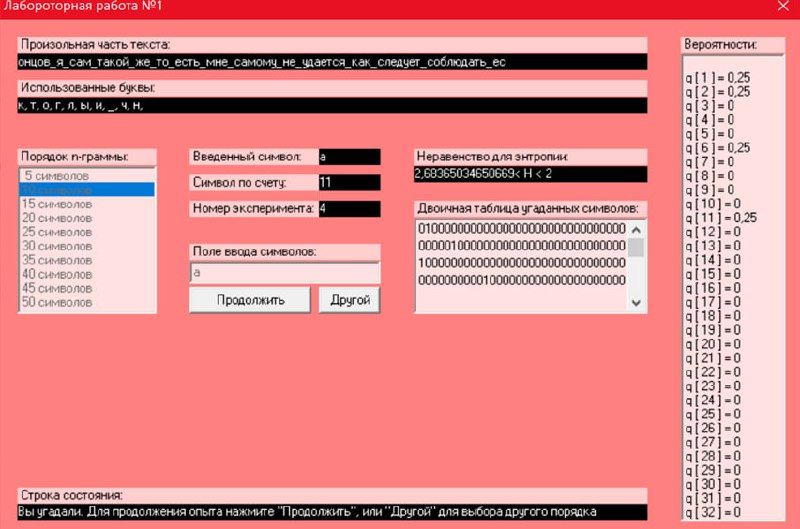
H∞​ — це ентропія «нескінченного порядку» (умовна ентропія, яку оцінюють через експеримент з n-грамами, наприклад H(10),H(20),H(30)

H0​=log2​m, де m — кількість символів в алфавіті (для російської мови без пробілу ≈ 32, з пробілом ≈ 33).

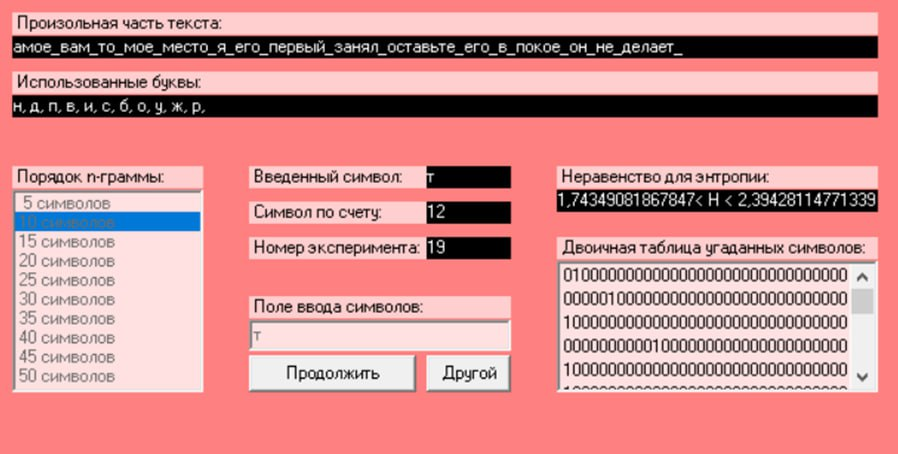
H(10)

Ентропія для H10: 2.028< H < 2.894

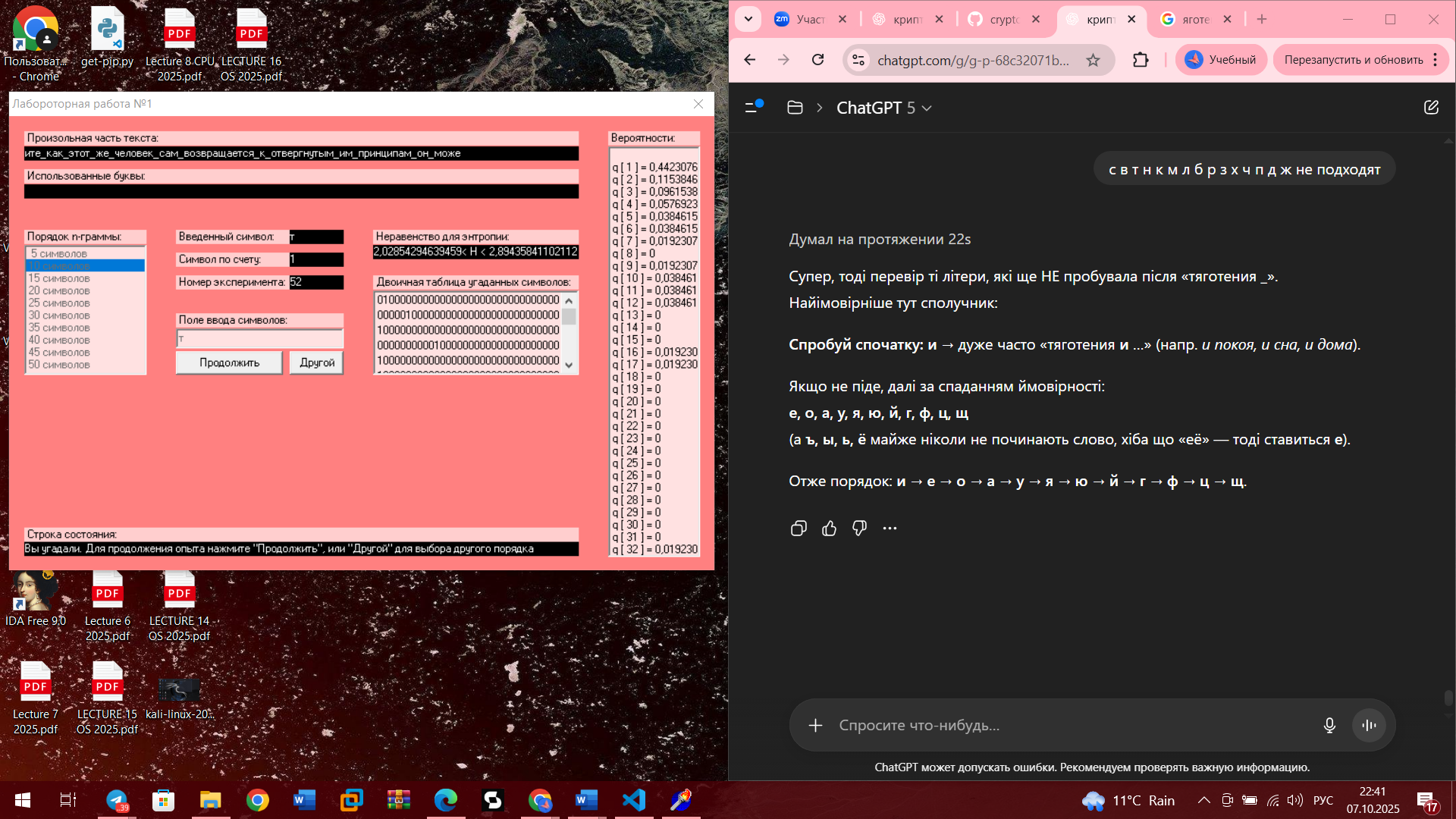
Надлишковість (R) 0.426< R <0.598



тут ми багато разів вгадували що може бути після с…



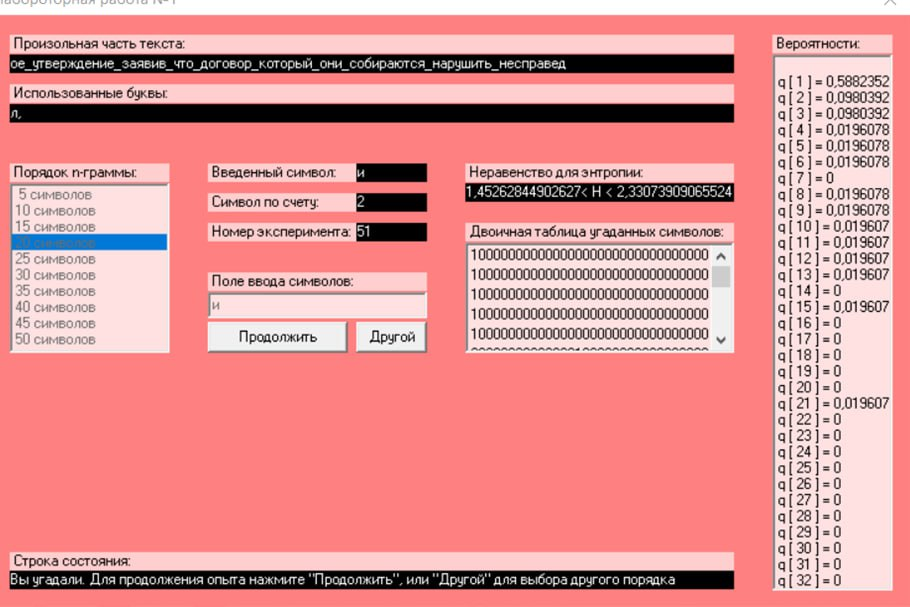
Найважче було вгадати перше слово, бо там було «амое вам\_» і треба було вгадати, що буде після пробілу



H(20)

Ентропія 1.453 < Н < 2.331

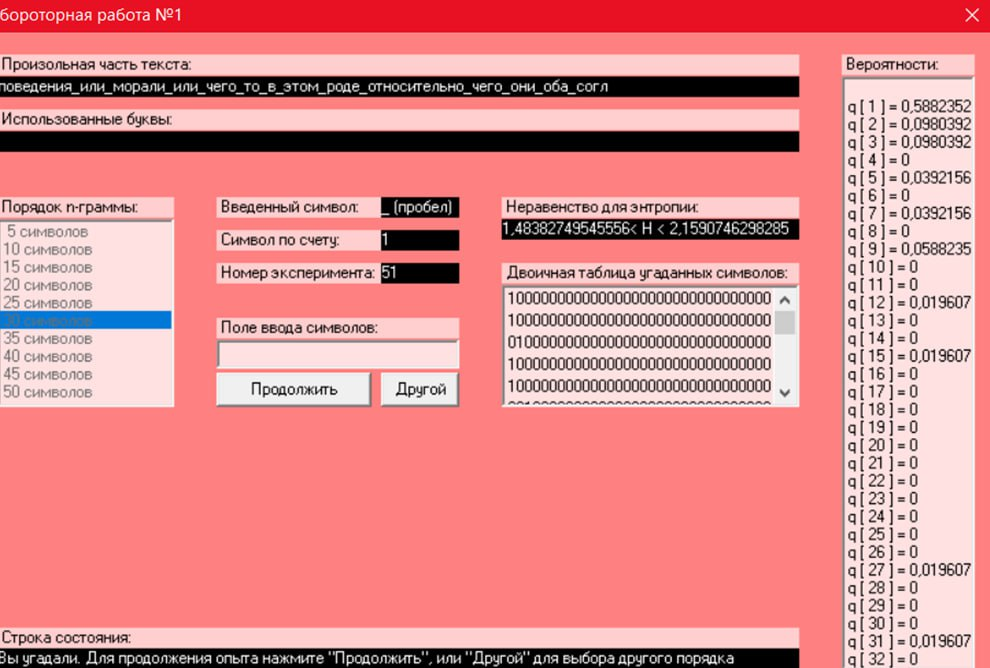
Надлишковість (R) 0.538< R <0.712



H(30)

Ентропія 1.484 < Н < 2.159

Надлишковість (R) 0.572< R <0.706



| **Порядок n-грам** | **Ентропія (H)** | **Надлишковість (R)** |
| --- | --- | --- |
| 10 символів | 2.028<𝐻<2.894 | 0.426<𝑅<0.598 |
| 20 символів | 1.453<𝐻<2.331 | 0.538<𝑅<0.712 |
| 30 символів | 1.484<𝐻<2.159 | 0.572<𝑅<0.706 |

**Висновок:**

У роботі на основі книги російською мовою підраховано частоти літер і біграм для двох варіантів (з пробілом/без пробілу) та побудовано матриці біграм з перетином і без, що загалом дало 7 таблиць результатів.

За програмою CoolPinkProgram було досліджено ентропію та надлишковість відкритого тексту за допомогою n-грамного аналізу. Метою було визначити міру невпорядкованості тексту (ентропію 𝐻) та частку повторюваної інформації (надлишковість 𝑅).

Отримані результати:

𝐻10:2.028<𝐻<2.894

𝑅10: 0.426<𝑅<0.598

𝐻20:1.452<𝐻<2.330

𝑅20: 0.538<𝑅<0.712

𝐻30:1.483<𝐻<2.159

𝑅30: 0.572<𝑅<0.706

Аналіз показав, що зі збільшенням порядку n-грам ентропія зменшується, а надлишковість зростає.