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**

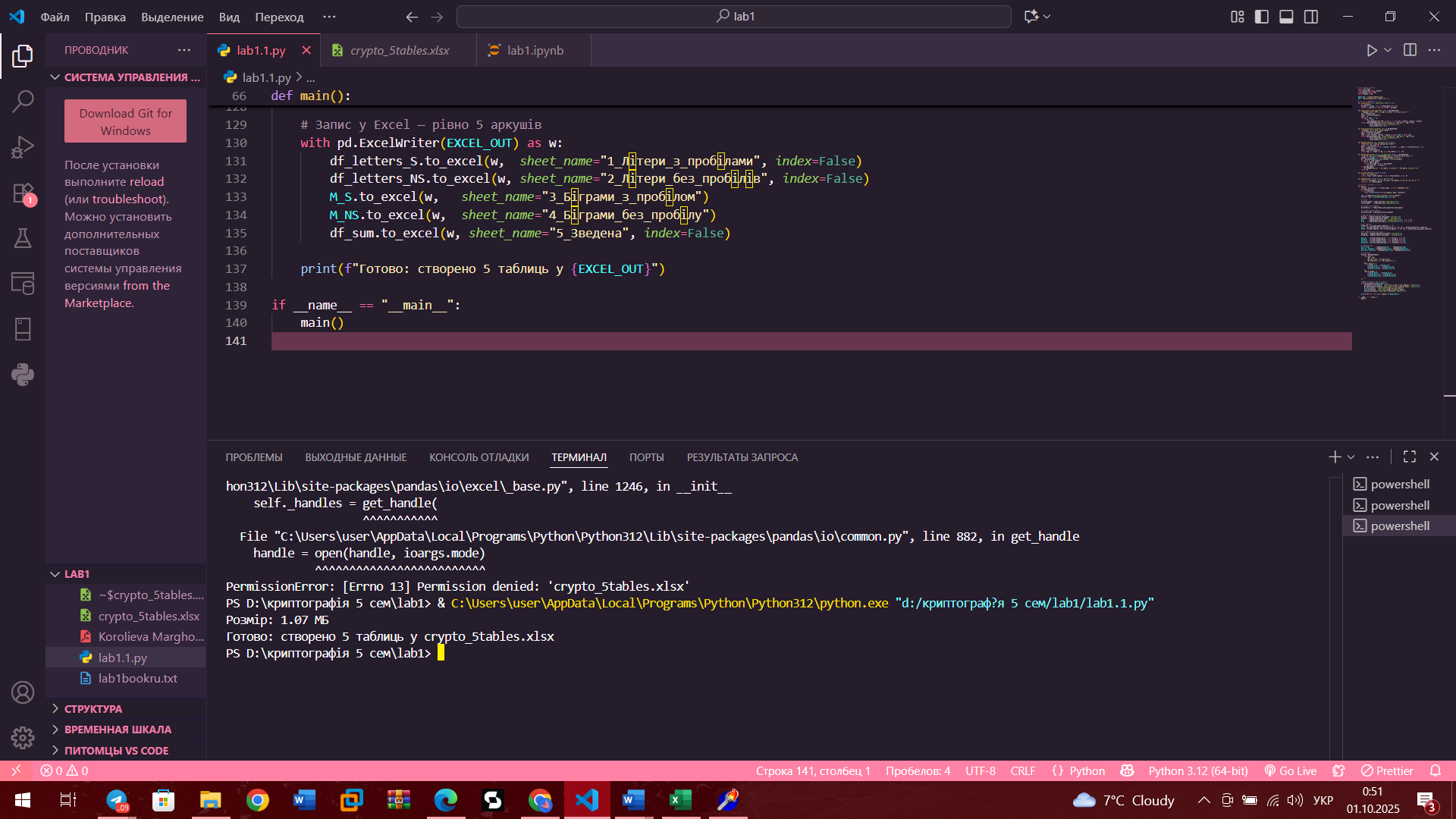
1. Написати програми для підрахунку a) частот букв і b) частот біграм в тексті, а також підрахунку c) H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

А) частота букв

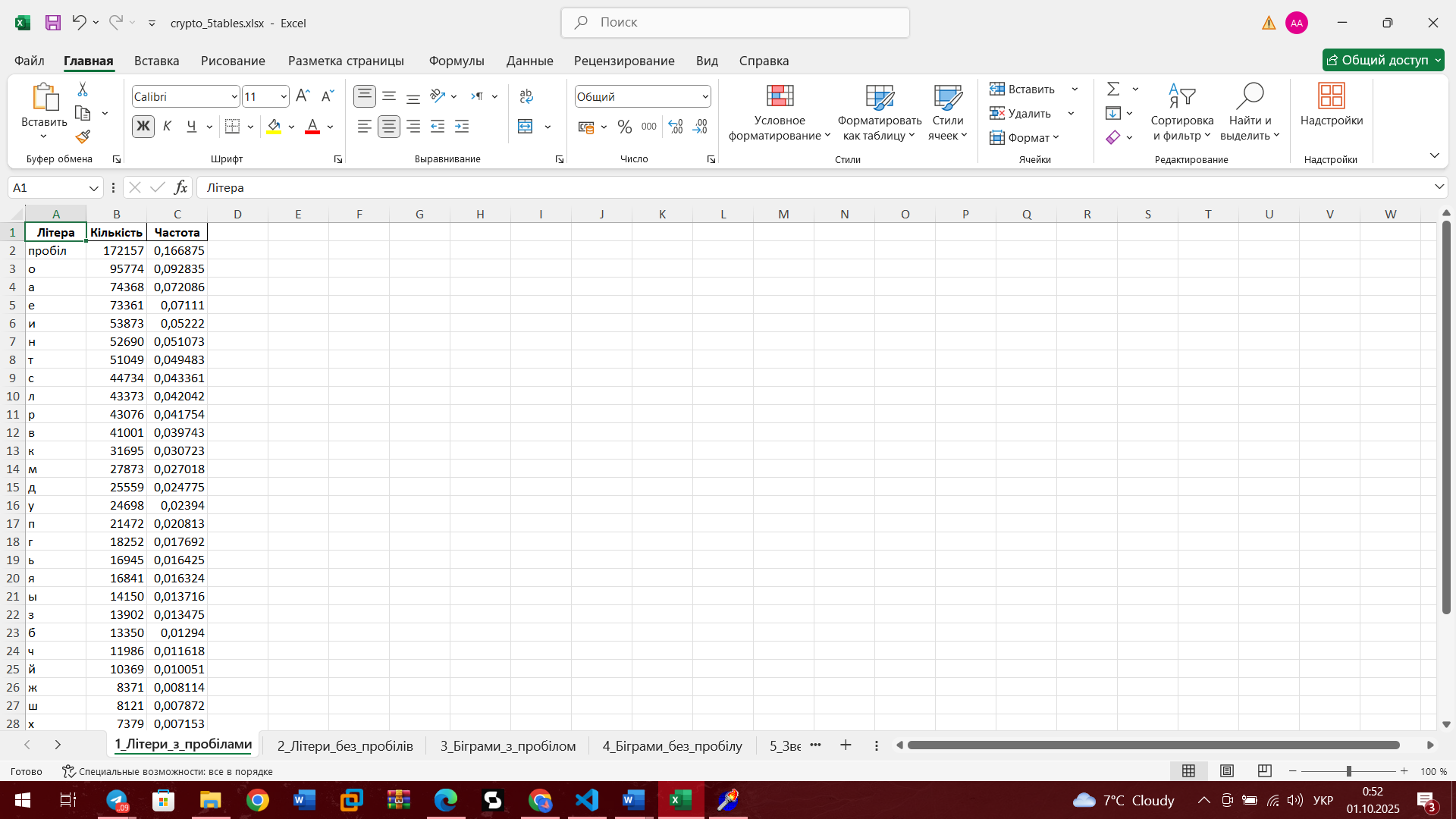
1. Частота літер

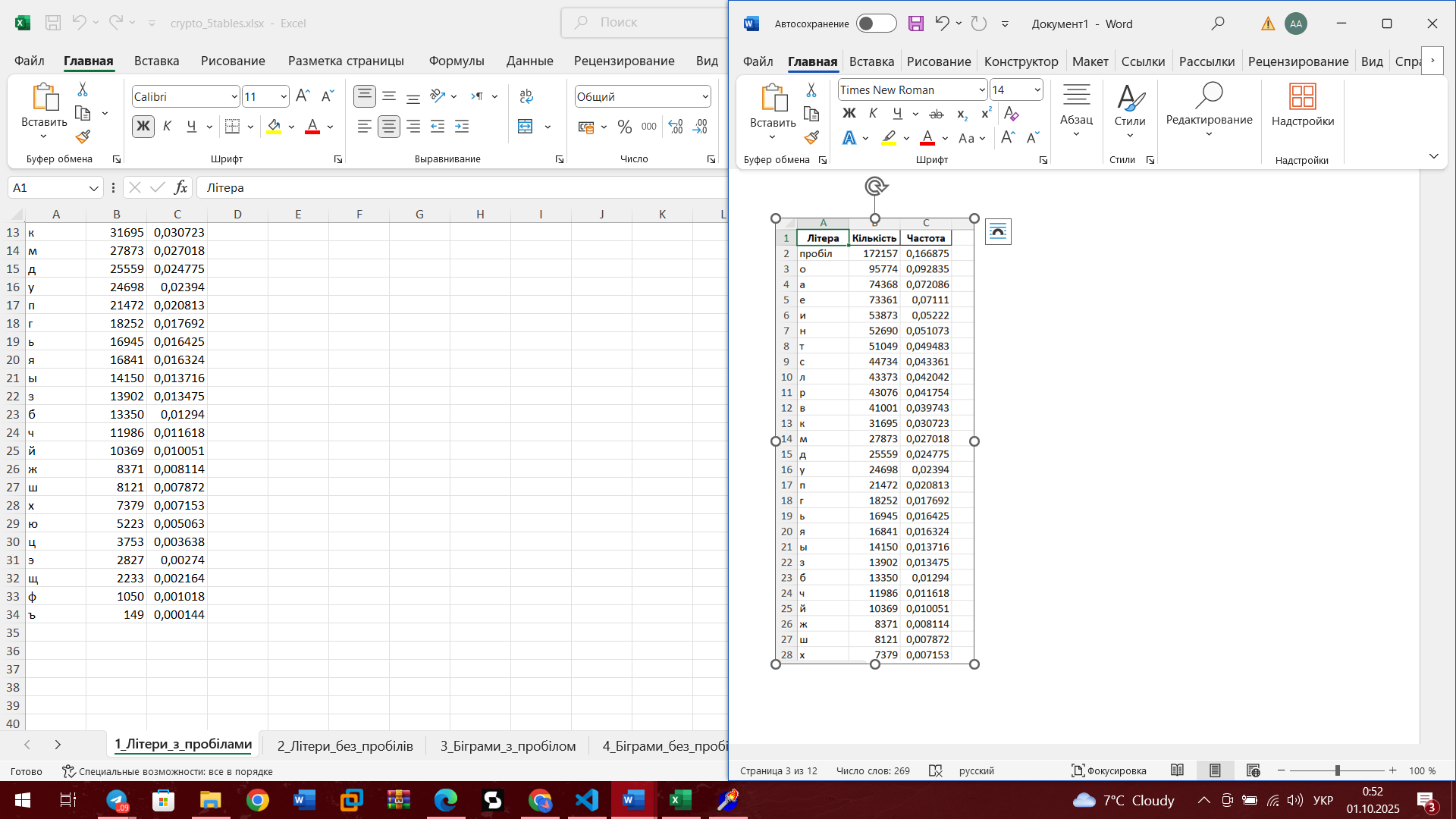
Частота окремої літери обчислюється як відношення кількості появ цієї літери до загальної кількості літер у тексті

𝑃(𝑙) =

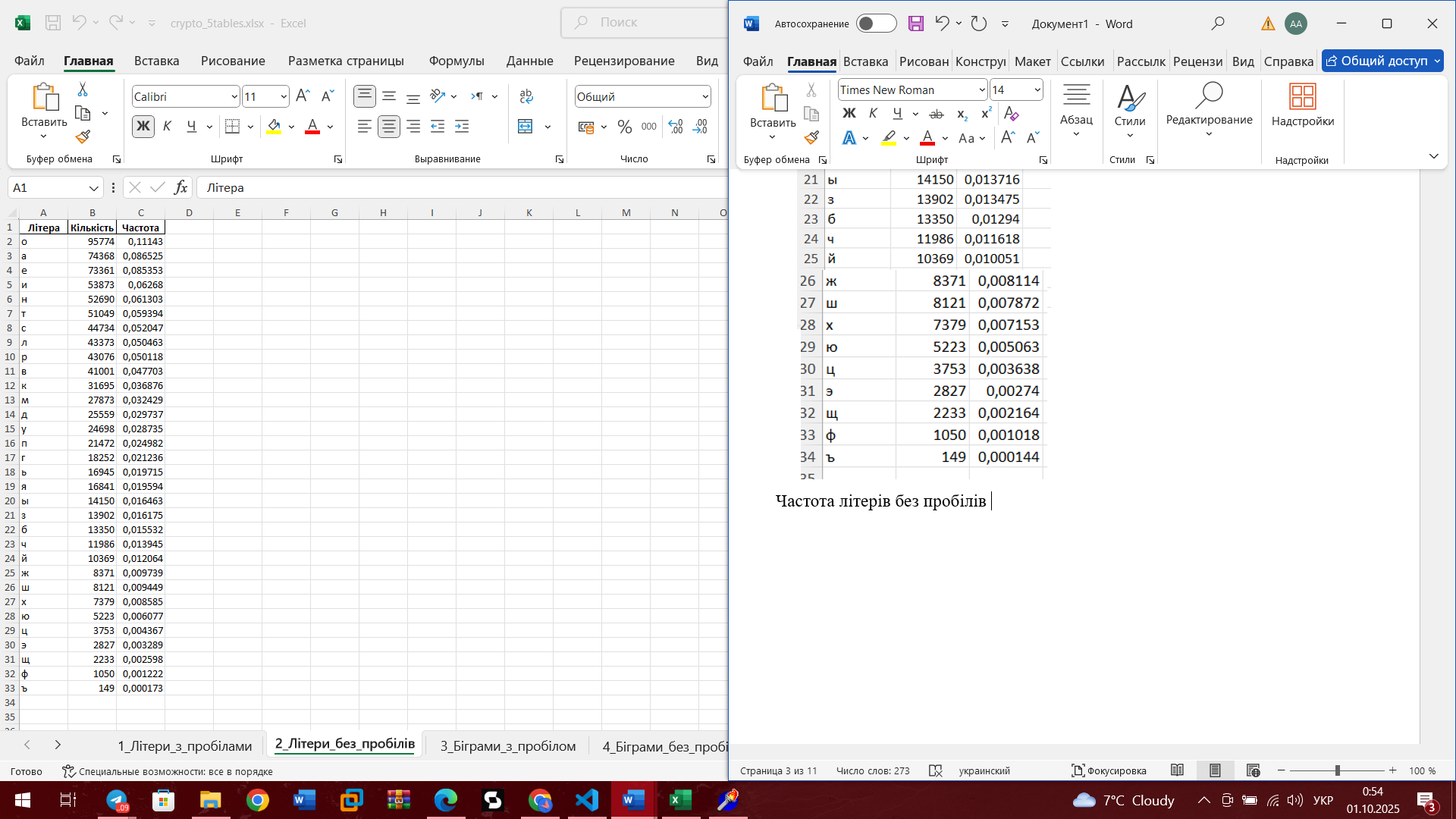
Дані збереглись а табличці

Частота літери з пробілами

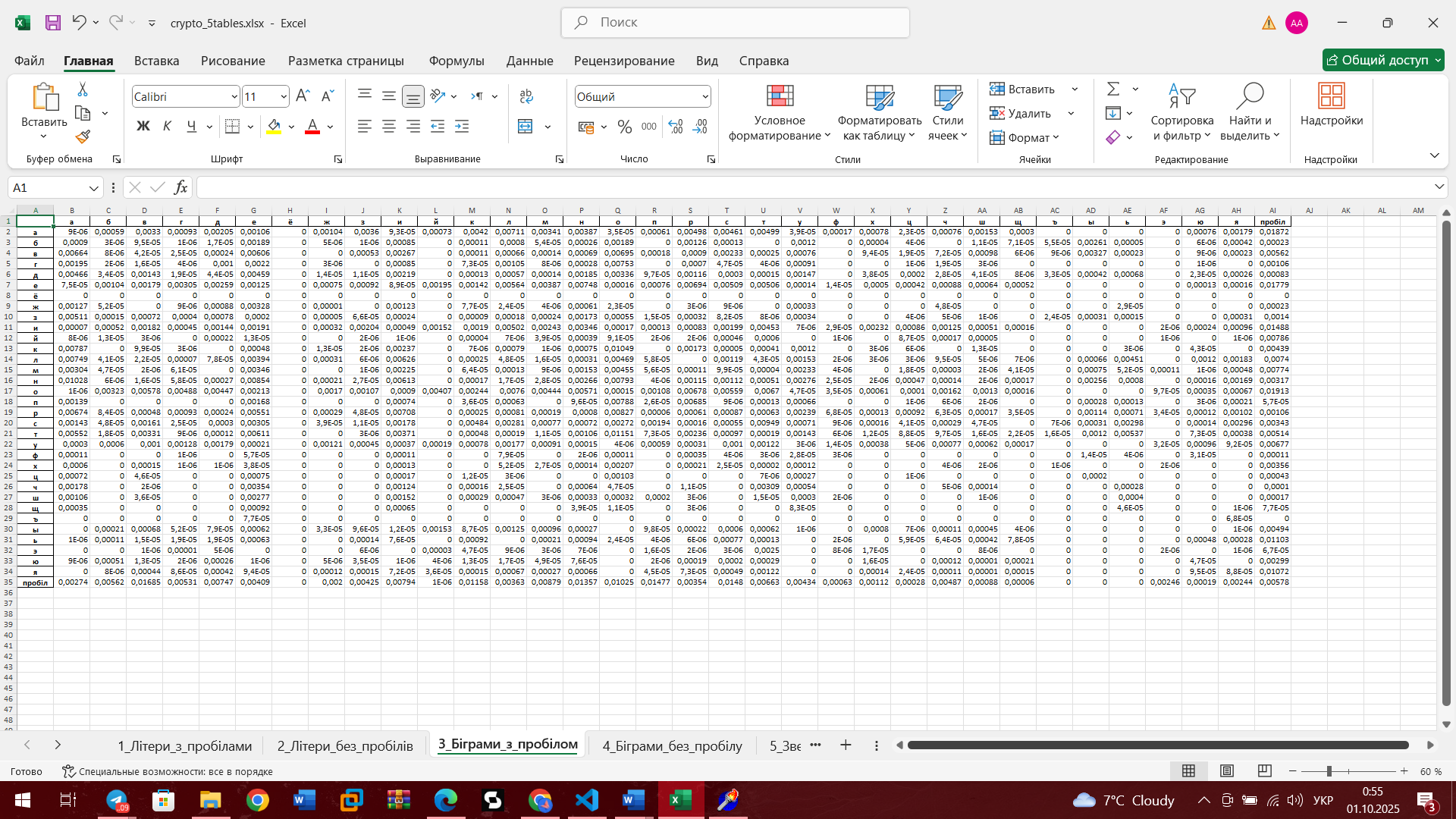
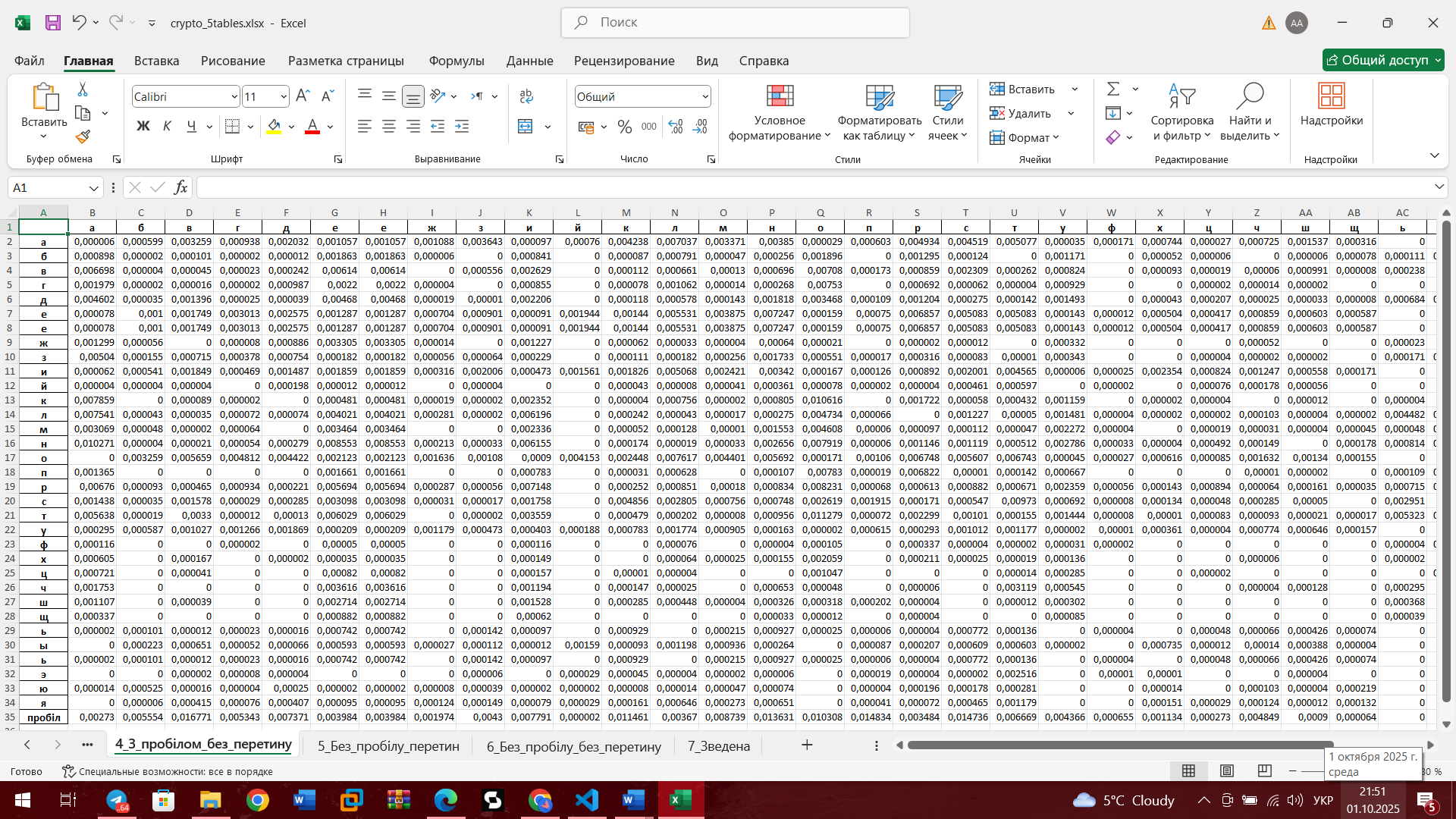


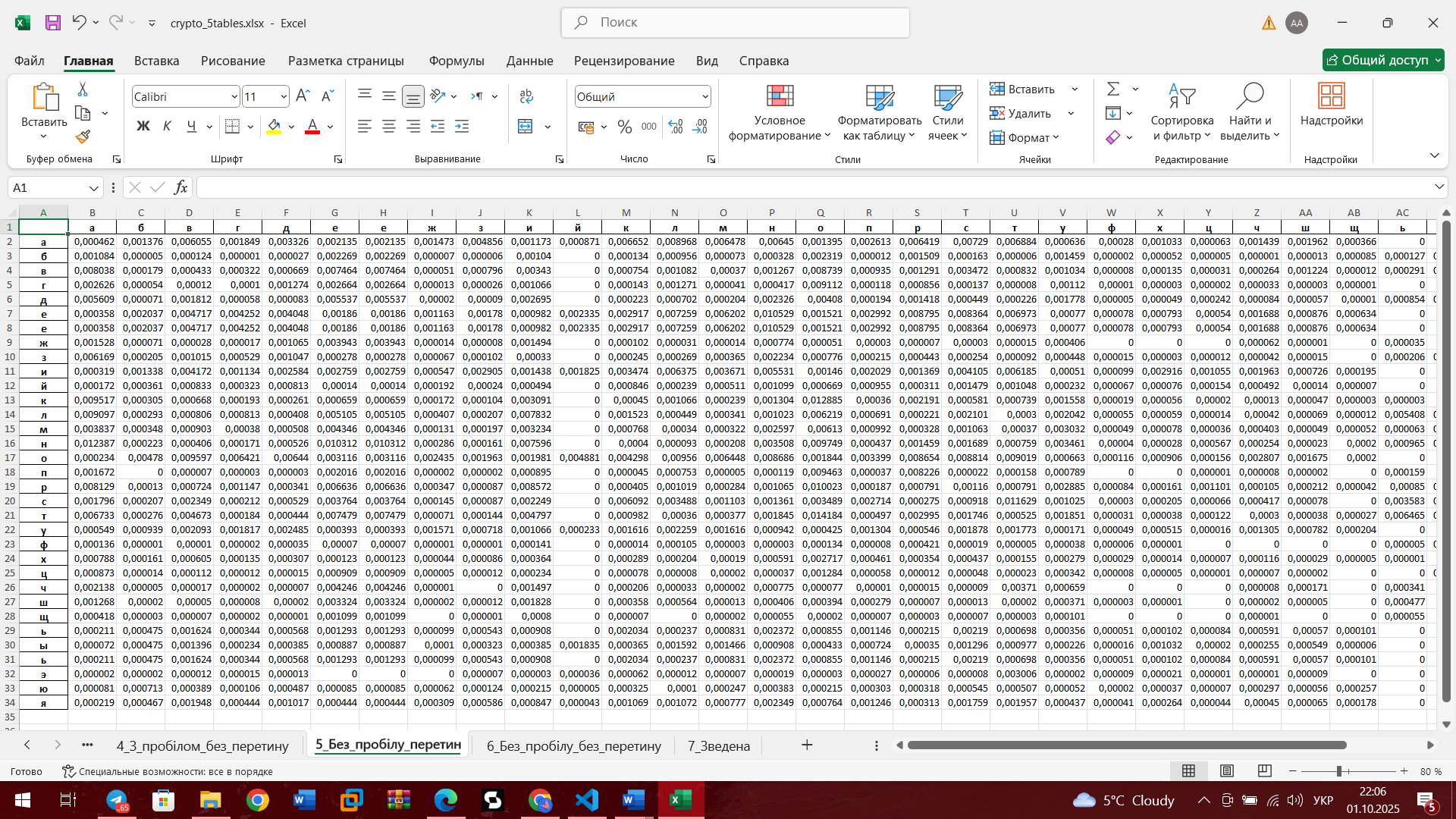


Частота літерів без пробілів

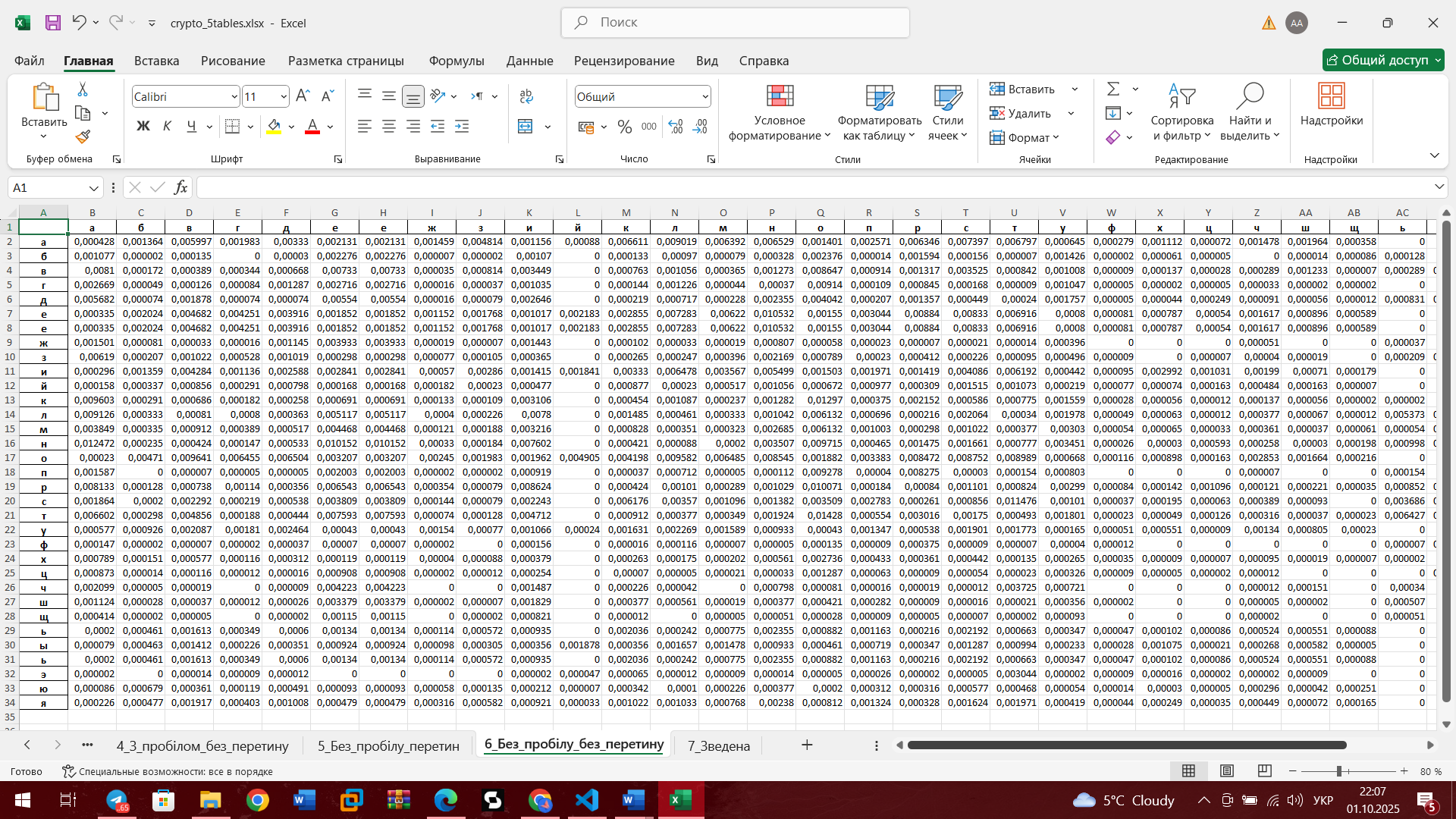


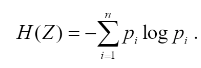
Біграми які перетинаются з пробілом

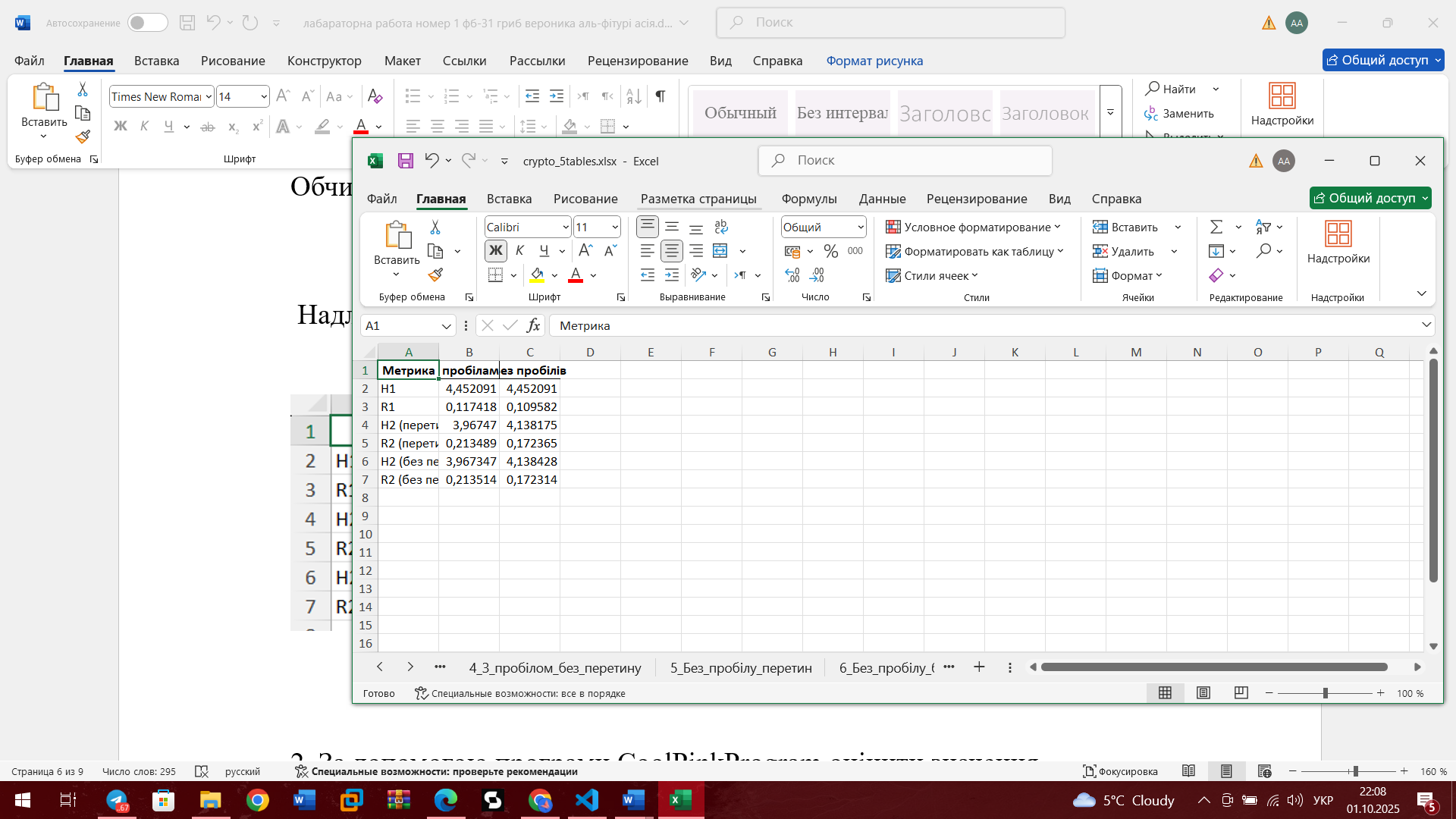
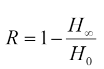
біграма з пробілу без перетену

без пробілу перетин

без пробілу без перетену

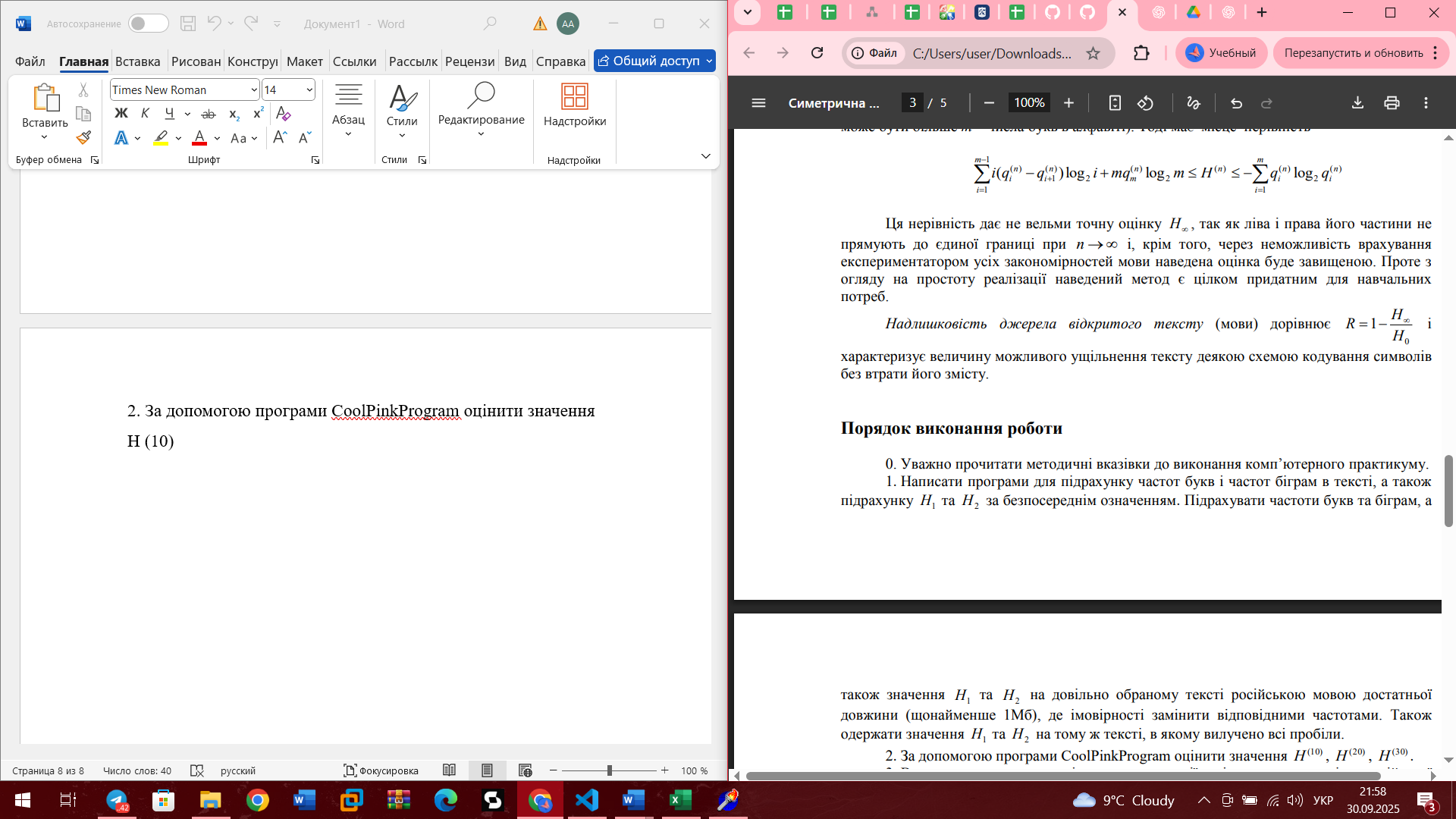


Обчислюємо ентропію символів за формулою: 

 Надлишковість: 

2. За допомогою програми CoolPinkProgram оцінити значення

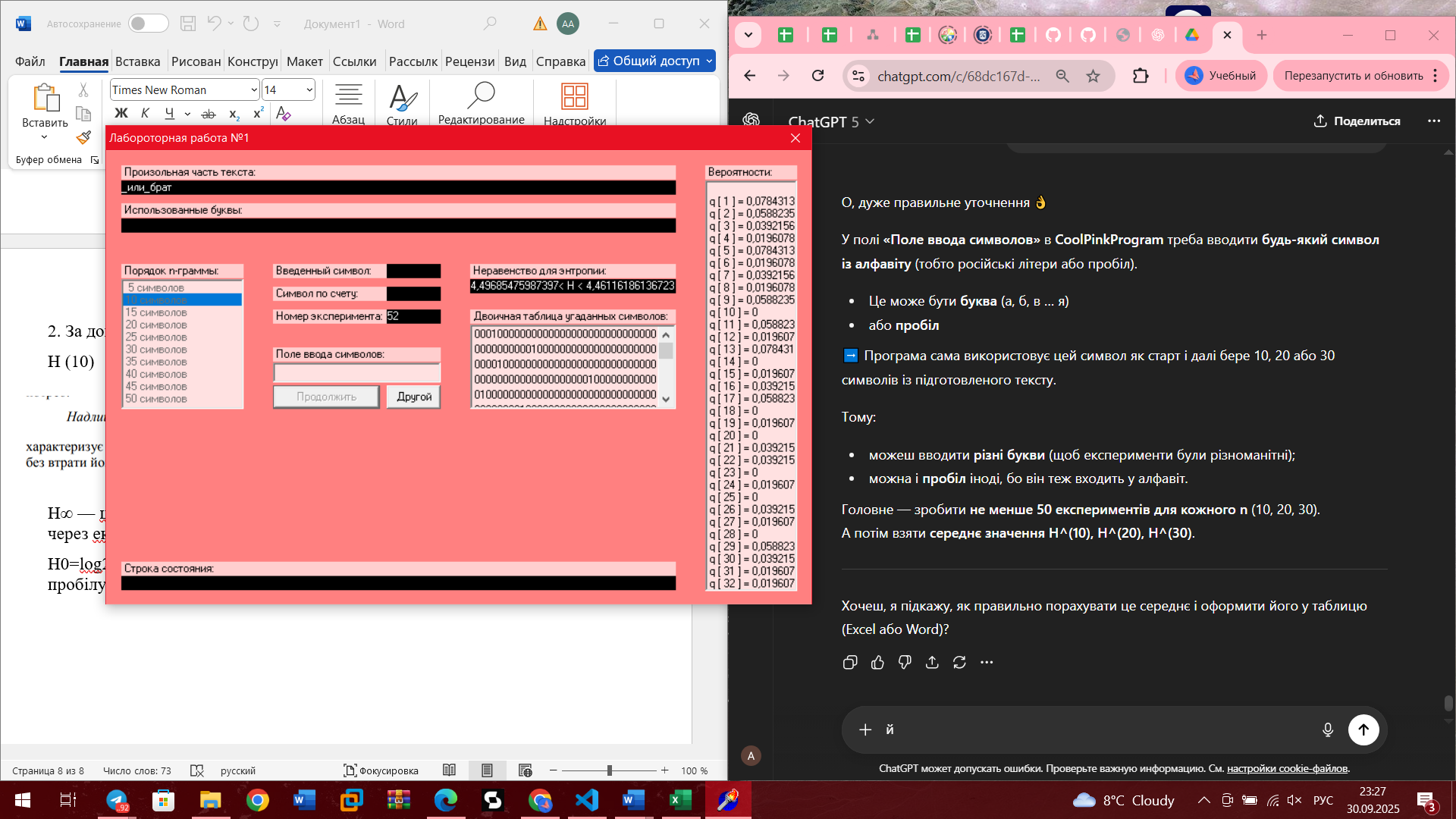
H (10), H (20) , H (30)



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

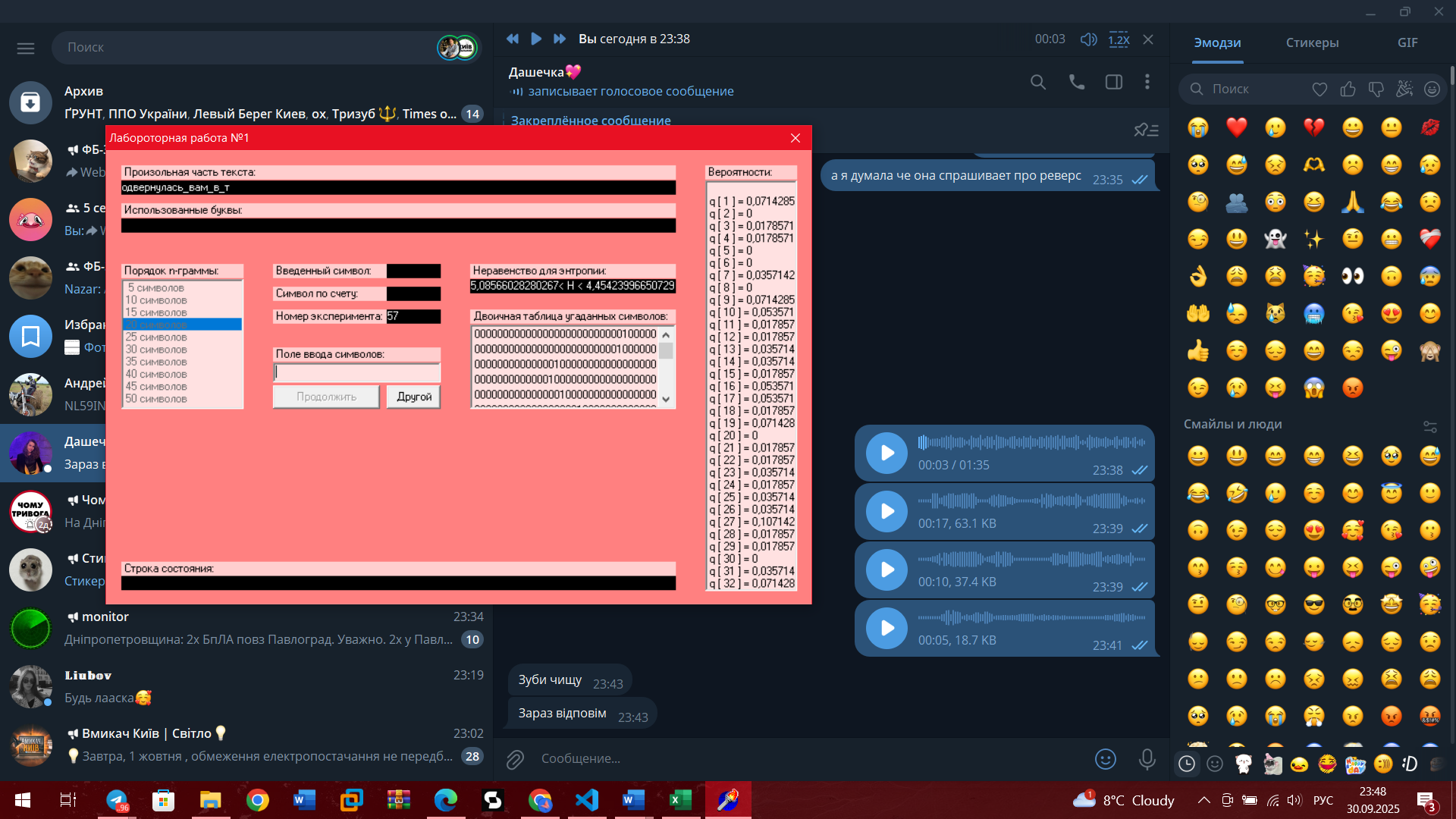
H0​=log2​m, де m — кількість символів в алфавіті (для російської мови без пробілу ≈ 32, з пробілом ≈ 33).

H(10)

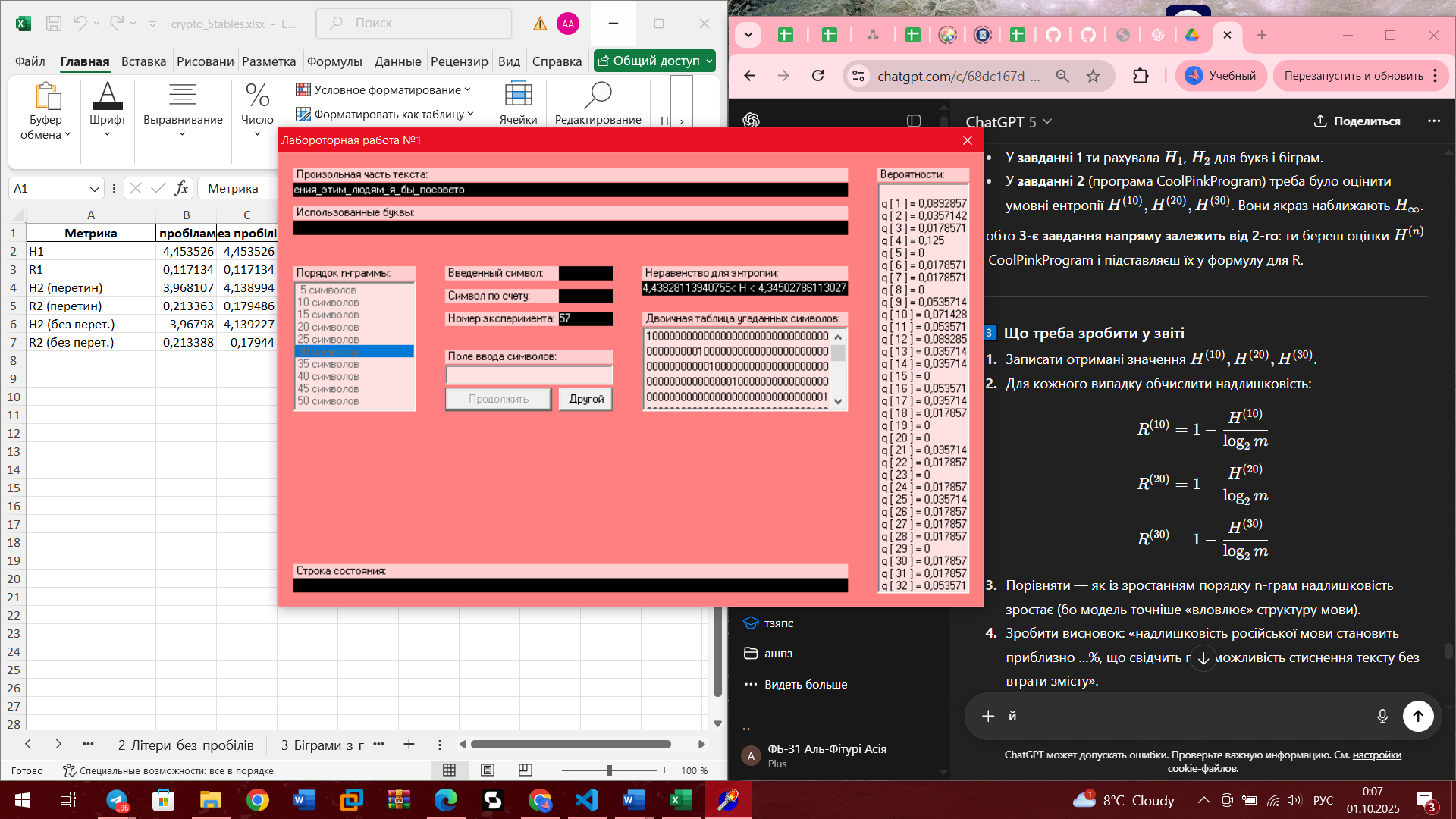
Ентропія 4,4968 < Н < 4,4611

H(20)

Ентропія 5,0856 < Н < 4,4542



H(30) Ентропія

4,4382 < Н < 4,3450

| **Порядок n-грам** | **Ентропія (H)** | **Надлишковість (R)** |
| --- | --- | --- |
| 10 символів | 4.4611 < H < 4.4968 | 0.110 < R < 0.116 |
| 20 символів | 4.4542 < H < 5.0856 | 0.000 < R < 0.117 |
| 30 символів | 4.3450 < H < 4.4382 | 0.120 < R < 0.138 |

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* отримано оцінки умовної ентропії H(n) та на їх основі обчислено надлишковість за формулою R=1−H\H0

H0= log2 33 це приблизно 5.044