МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. І. СІКОРСЬКОГО»

**Комп’ютерний практикум №1**

***Експериментальна оцінка ентропії на символ джерела відкритого тексту***

Виконав:

Студент групи ФБ-05

Левицький Євген

Київ - 2022

***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. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.
2. 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.
3. За допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30).
4. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

***Хід роботи***

1. Пишемо програму 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. Для коду використовуємо модулі: re – для модифікації тексту, math – отримати логарифм у формулі надлишковості, collections – використання Counter для підрахунків в тексті, pandas – запису в файл.

В результаті роботи було створено програму з функціями: очистки тексту, обчислення ентропії, частоту букв, підрахунок біграм, частота біграм, надлишковість.

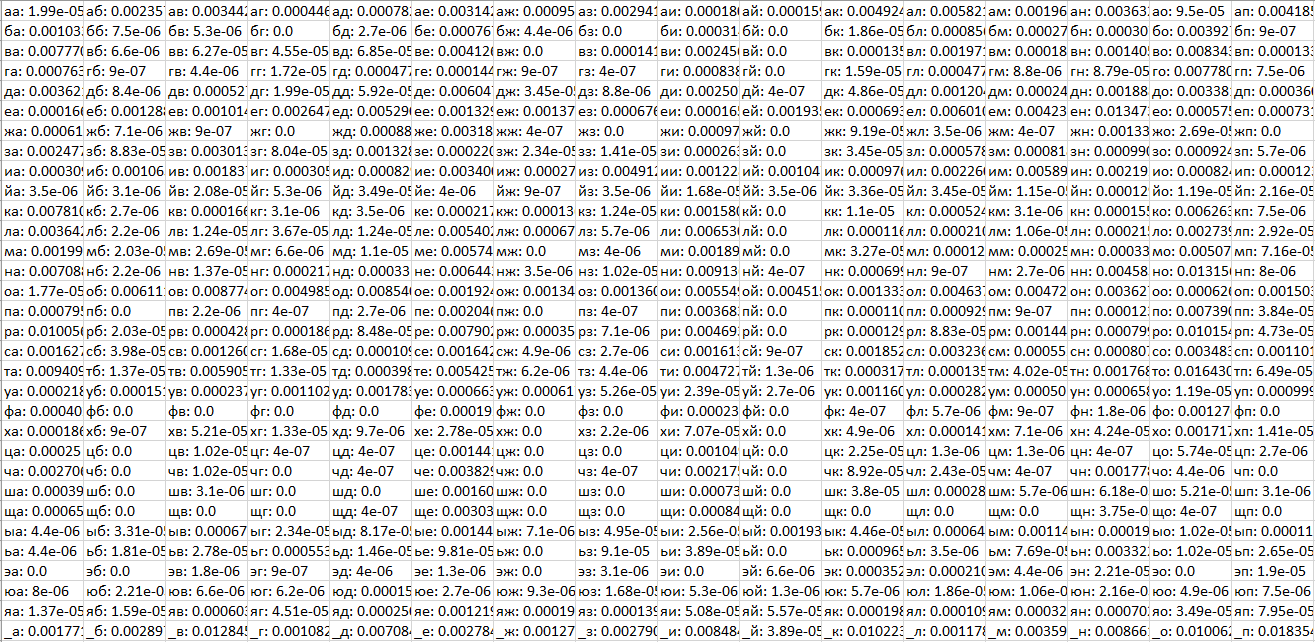
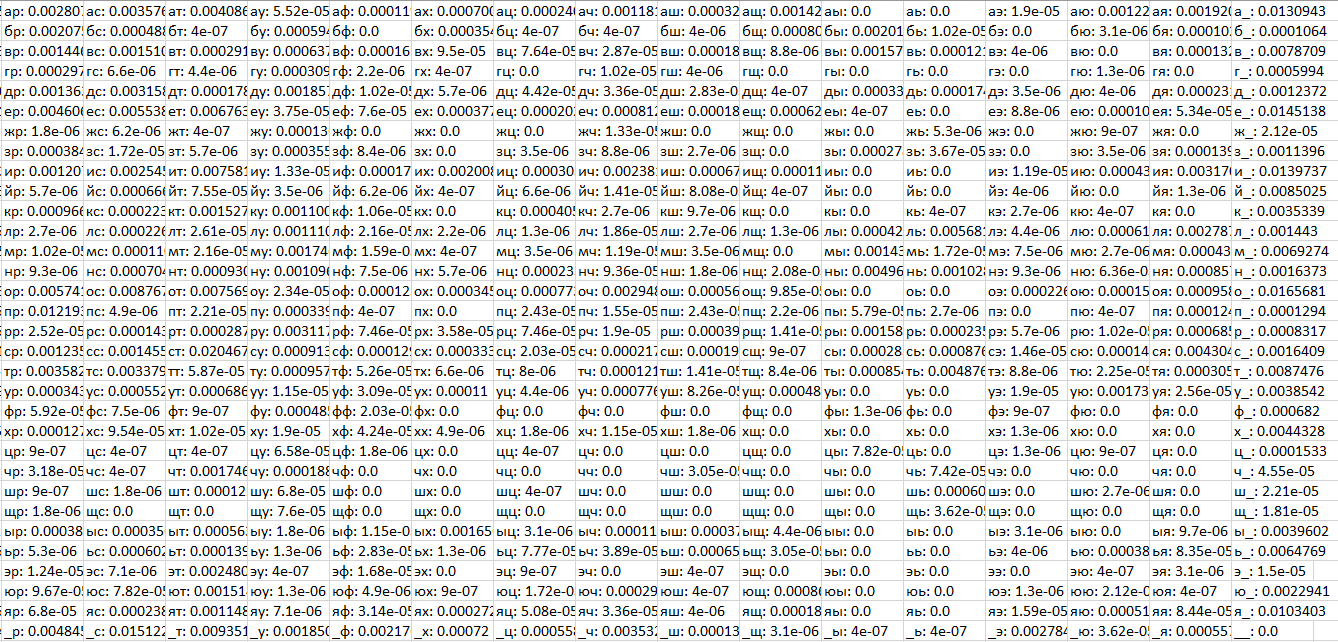
Частоти літер:

|  |  |
| --- | --- |
| Частота літер без пробілів | Частота літер з пробілами |
| 'о': 0.12073517880188848  'е': 0.08728487004394492  'т': 0.07821000032672798  'а': 0.07609494715174882  'и': 0.07176784343195072  'н': 0.06168266953098198  'с': 0.05731982993808505  'р': 0.050787822848087825  'в': 0.04738270465424011  'л': 0.03699837044418669  'д': 0.03310418130135755 'п': 0.032439494061718914  'м': 0.03046687386666231  'к': 0.02852692646986751  'у': 0.019858424681031807  'я': 0.019683829415320275  'б': 0.016395618577753093  'ы': 0.01604132292160162  'ь': 0.01593003120252234  'з': 0.015316905722640616  'ч': 0.014720116641889795  'г': 0.013709812458137977  'й': 0.011219022103148023  'ж': 0.008453065525296915  'х': 0.008135526767189975  'ю': 0.006351796187084443  'щ': 0.005432363223498276  'ш': 0.00462371146651855  'ф': 0.00389572068024766  'э': 0.003655269305539673  'ц': 0.003630254194370477  'ъ': 0.0001357963177756359  'ё': 9.699736983973992e-06 | ' ': 0.14712422372958991  'о': 0.1029721093438074  'е': 0.07444315129539138  'т': 0.06670341474076716  'а': 0.0648995371223036  'и': 0.06120905517827823  'н': 0.05260765465866743  'с': 0.04888669445413218  'р': 0.04331570383664697  'в': 0.04041156101377661  'л': 0.03155501391332592  'д': 0.02823375432519172  'п': 0.02766685867970788  'м': 0.02598445869956229  'к': 0.024329924557597162  'у': 0.016936769365342472  'я': 0.016787861292565612  'б': 0.013983425921934725  'ы': 0.013681255739165305  'ь': 0.013586337727863095  'з': 0.013063417858257802  'ч': 0.01255443090774274  'г': 0.011692766942756167  'й': 0.00956843218521726  'ж': 0.007209414821752248  'х': 0.006938593706935851  'ю': 0.005417293103771075  'щ': 0.00463313100122392  'ш': 0.003943451506257405  'ф': 0.003322565799298913  'э': 0.0031174906464395513  'ц': 0.0030961558640826324  'ъ': 0.00011581738993755876  'ё': 8.272670709825625e-06 |

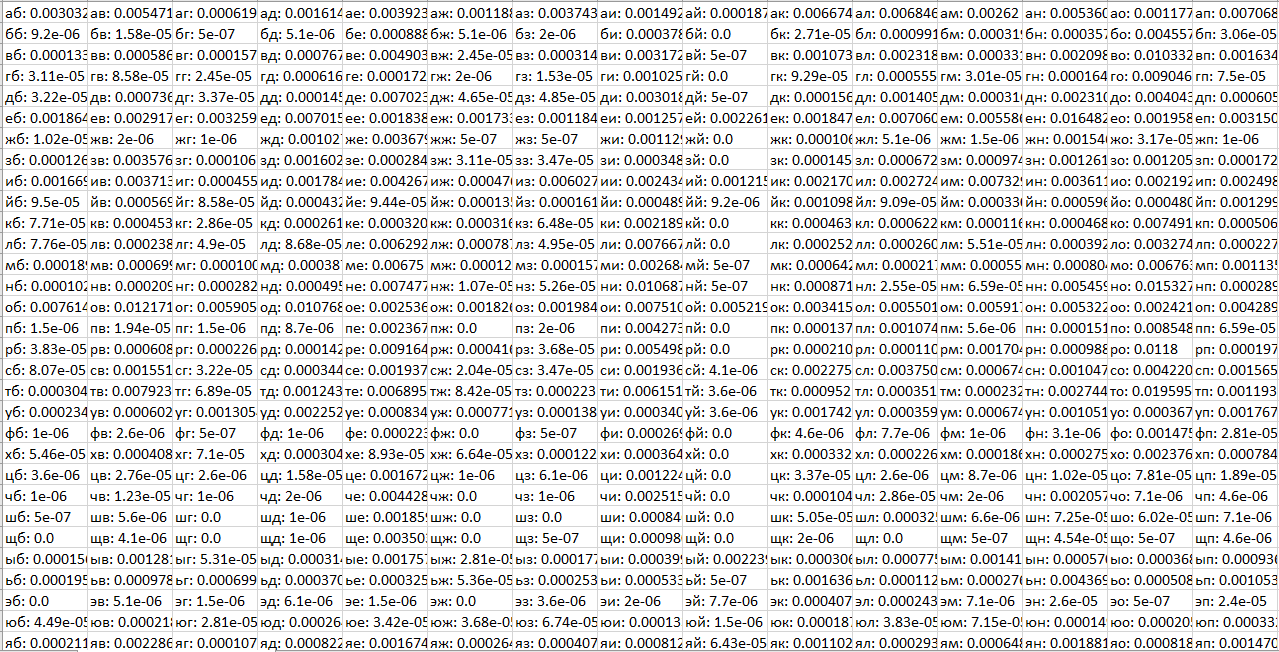
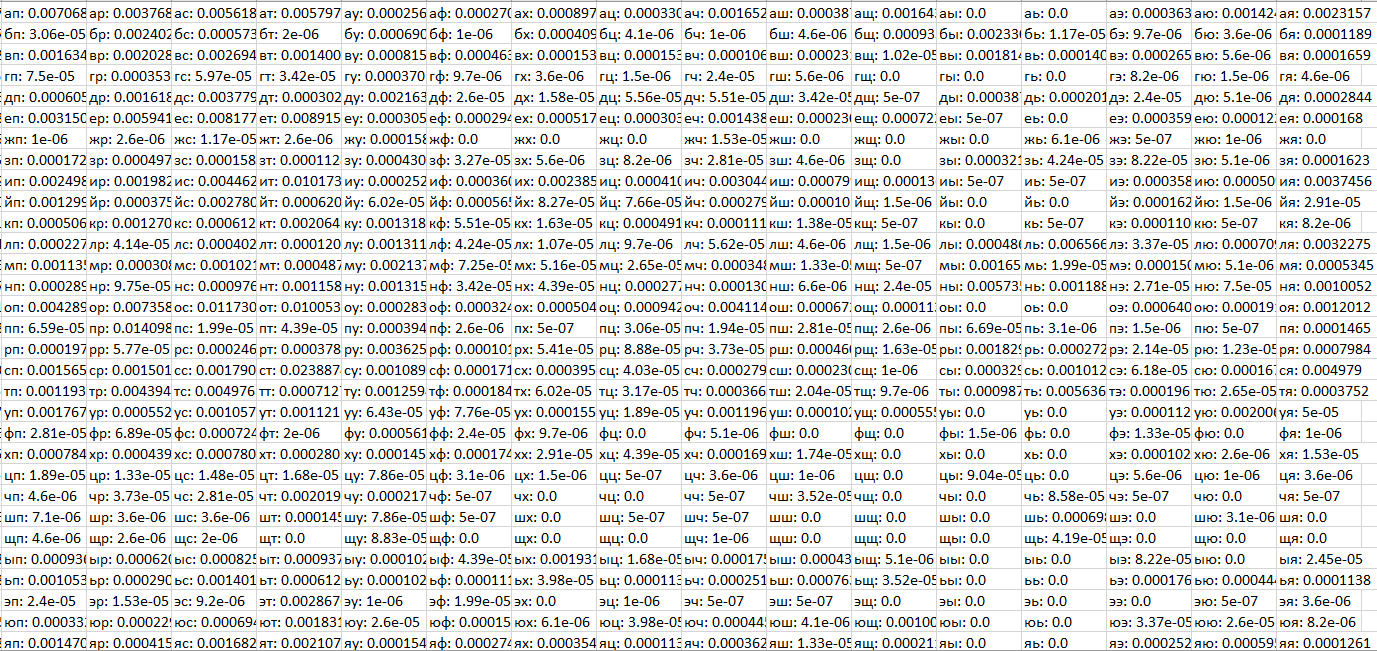
Ентропія монограм з пробілами: H1 = 4.3635721579, R = 0.11921694

Ентропія монограм без пробілів: H1 = 4.4097577006, R = 0.10989443

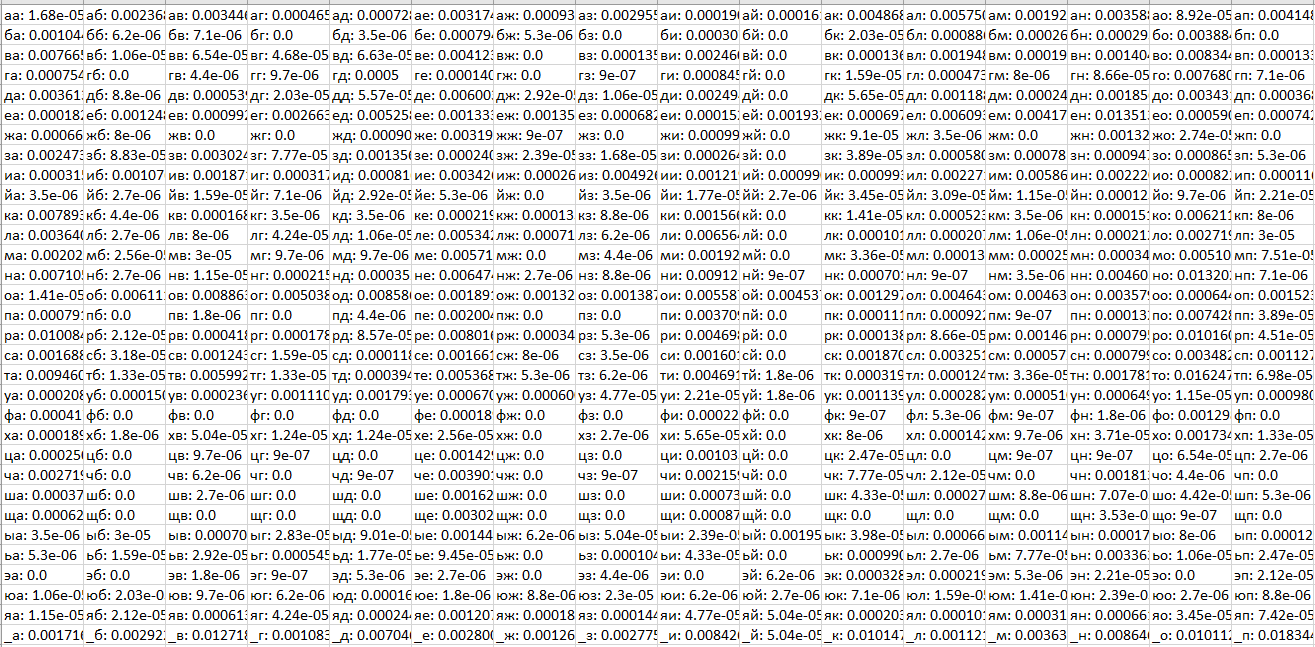
Біграм перехресна з пробілами

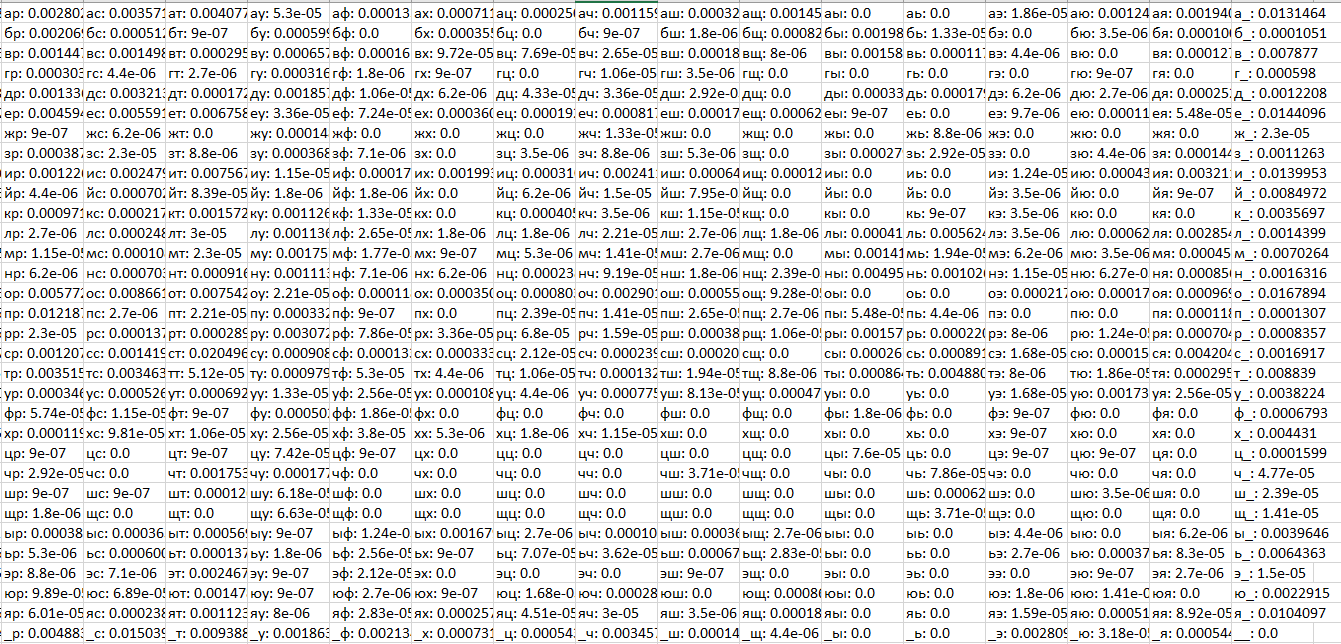
  


H = 4.385750539048615 R = 0.11474025971608492

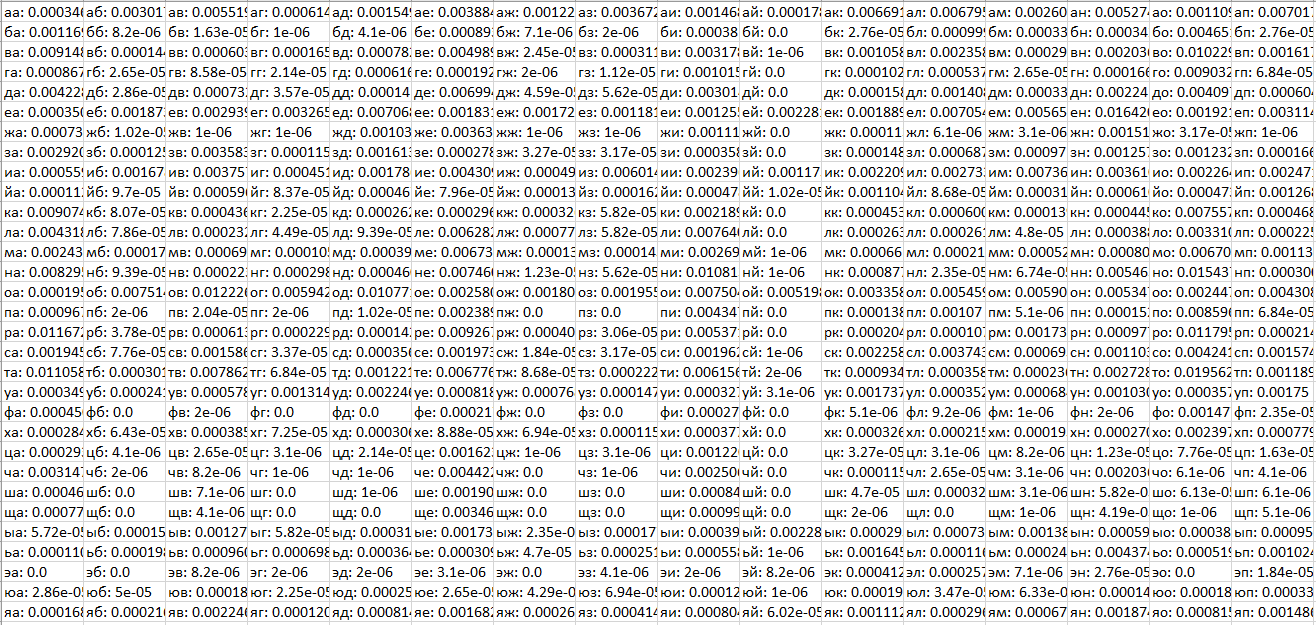
Біграм перехресна без пробілів  
  
  
H = 4.0269650328813595 R = 0.1871607864148418

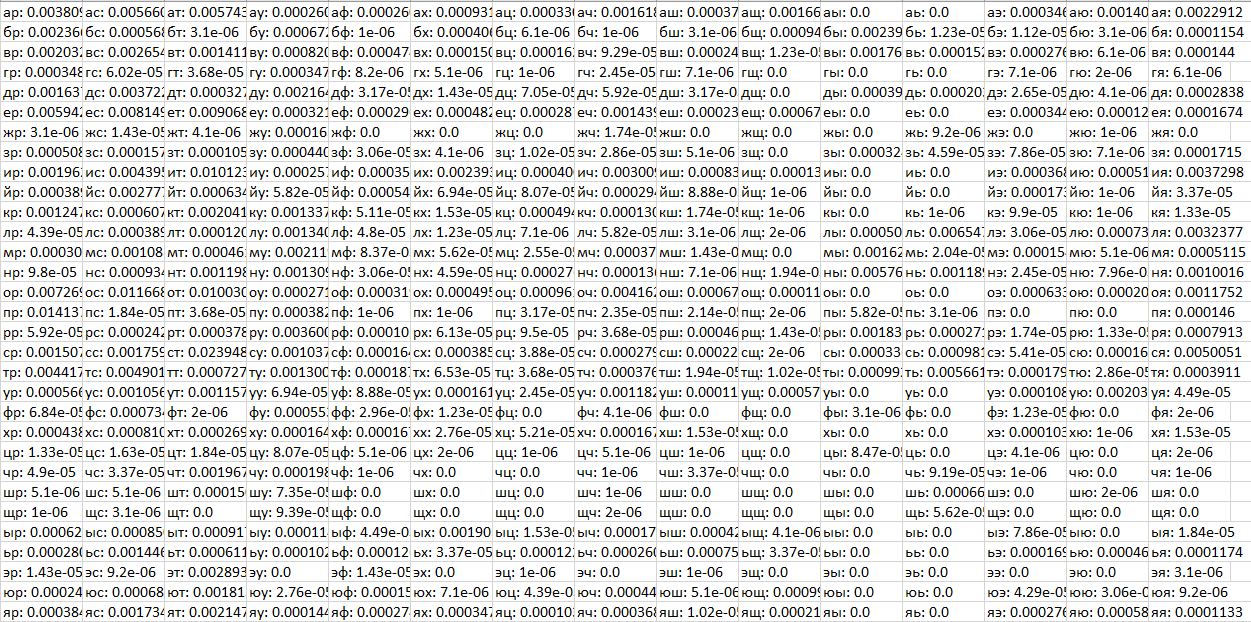
Не перехресна з пробілами

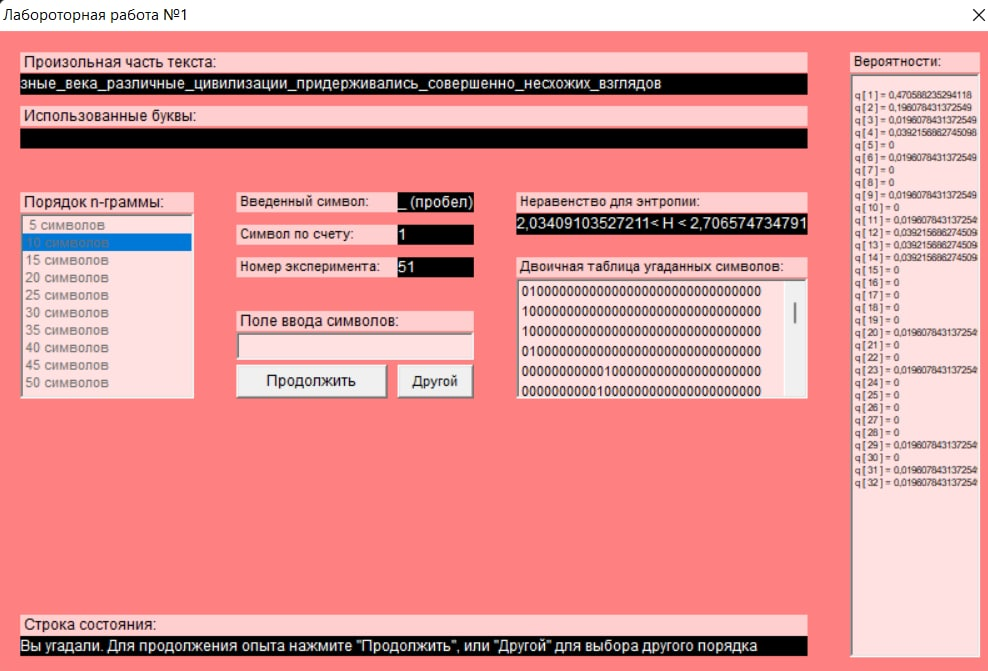
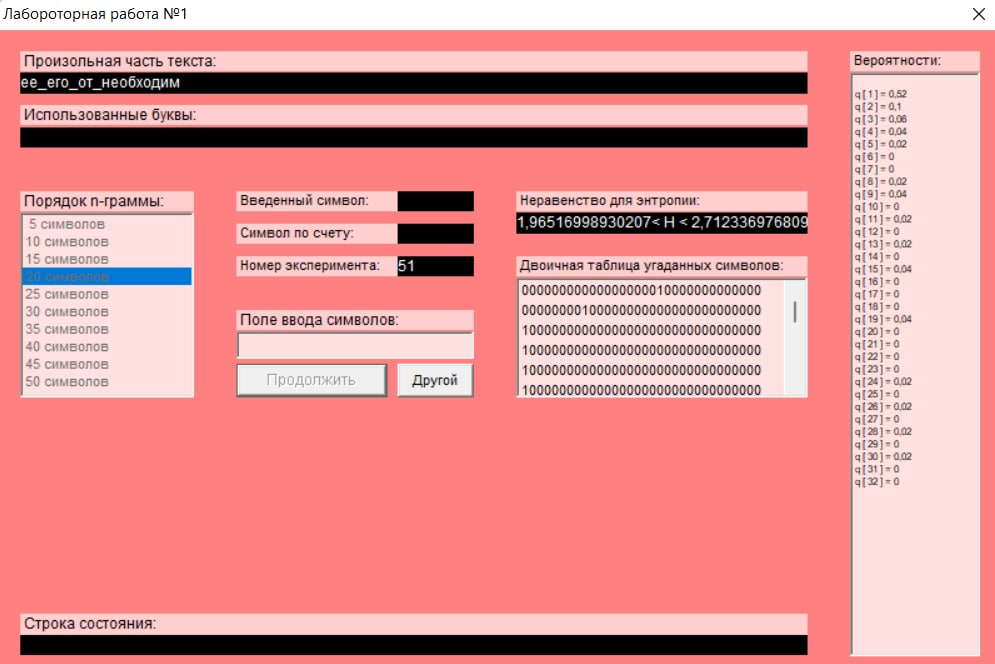
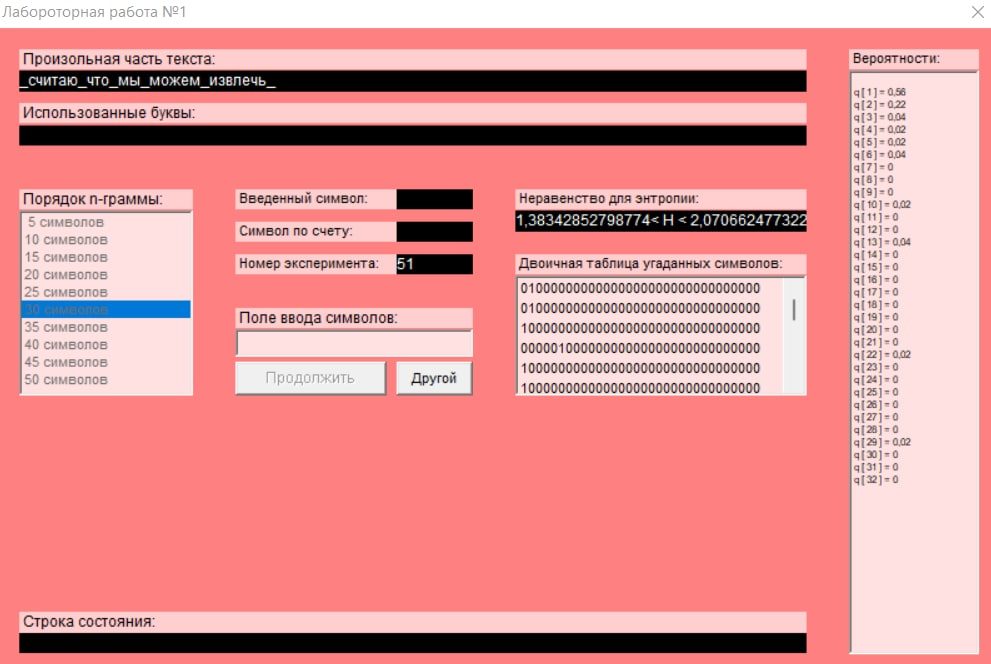


  
H = 3.92460193346591 R = 0.20782268453156305

Не перехресна без пробілів



  
H = 4.026597331888821 R = 0.187235006524321

Робота з CoolPinkProgram  
  
H10: 2,03409 < H < 2,70657  
R: 0,59318 < R < 0,458686   
  
H20:1,96516 < H < 2,71233  
R: 0,60696 < R < 0,45753  
  
  
  
  


H30: 1,38342 < H < 2,07066  
R: 0,72331 < R < 0,585868

***Висновки***

В процесі виконання практикуму були набуті практичні навички щодо обчислення ентропії,надлишковості, частот літер і біграм. Ознайомились з поняттями ентропії та надлишковості, отримали досвід отримання умовної ентропії на прикладі програми CoolPinkProgram