НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ» ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

Лабораторна робота №1

Виконали:
студенти ФІ-04

Кравченко Антон

Давидюк Данил

Київ – 2023

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:

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
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:

Для роботи було обрано декілька глав з твору "Мертві душі"
Так як роки написання цього твору припадають на 1842, то алфавіт дещо відрізняється від теперішнього поняття алфавіту (входять більше літер).

Результат роботи програми за посиланням якщо пробіл **не входить** до алфавіту
Результат роботи програми за посиланням якщо пробіл **входить** до алфавіту

1.1 Кількість та частота літер, якщо вважати, що пробіл не входить до алфавіту:

'r': 2	'r': 4.1968578125557394e-6
'a': 37185	'a': 0.0780300788799426
'o': 53516	'o': 0.11229952134836647
'щ': 1399	'щ': 0.00293570203988274
'o': 2	'o': 4.1968578125557394e-6

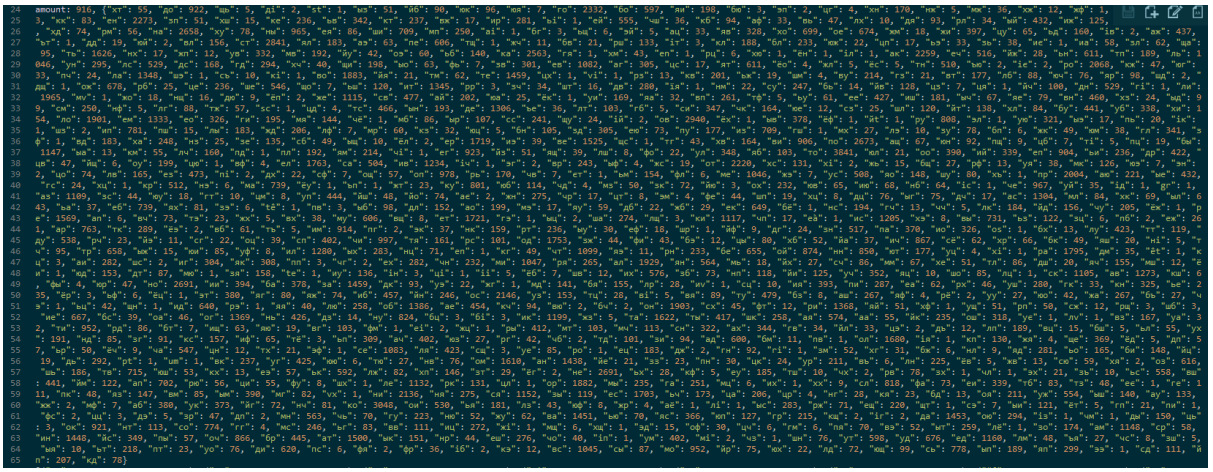
'м': 14218	'м': 0.02983546218945875
'і': 73	'і': 0.0001531853101582845
'п': 13525	'п': 0.028381250957408188
'm': 1	'm': 2.0984289062778697e-6
'д': 14922	'д': 0.03131275613947837
'э': 965	'э': 0.0020249838945581443
'ю': 3144	'ю': 0.006597460481337623
'ы': 9142	'ы': 0.019183837061192286
'л': 21292	'л: 0.0446797482724684
't': 6	't': 1.2590573437667218e-5
'u': 1	'u': 2.0984289062778697e-6
'à': 1	'à': 2.0984289062778697e-6
'є': 8	'є': 1.6787431250222958e-5
'ц': 1761	'ц': 0.0036953333039553287
'ъ': 138	'ъ': 0.000289583189066346
'v': 5	'v': 1.049214453138935e-5
'в': 21803	'в': 0.045752045443576396
'н': 29307	'н': 0.06149865595628553
'й': 4971	'й': 0.01043129009310729
'у': 14621	'у': 0.030681129038688734
'g': 1	'g': 2.0984289062778697e-6
's': 3	's': 6.295286718833609e-6
'г': 8523	'г': 0.017884909568206285
'з': 7996	'з': 0.016779037534597845
'к': 20107	'к': 0.04219311001852913
'ч': 8888	'ч': 0.018650836118997708
'e': 40316	'e': 0.0846002597854986
'б': 9176	'б': 0.019255183644005733
'ь': 9568	‘ь’ 0.020077767775266656
'и': 32585	'и': 0.06837730591106439
'х': 4798	'х': 8.393715625111479e-6
'р': 20000	'р': 0.0419685781255574
'e': 3	'e': 6.295286718833609e-6
'ê': 2	'ê': 4.1968578125557394e-6
'я': 9210	'я': 0.01932653022681918

'х': 4	'х': 0.010068261892321219
'ф': 503	'ф': 0.0010555097398577684
'ё': 130	'ё': 0.00027279575781612305
'n': 14	'n': 2.9378004687890178e-5
'с':1	'с': 2.0984289062778697e-6
'ш': 4840	'ш': 0.01015639590638489
'ж': 5172	'ж': 0.010853074303269143
'ї': 1	'ї': 2.0984289062778697e-6
'р': 2	'р': 4.1968578125557394e-6
'с': 24014	'с': 0.05039167175535676

H1 chars = 4.479795715390369
H2 chars = 2.2398978576951847

1.2 Кількість та частота біграм без перетенів без пробілів:

Кількість: строка 24-65.



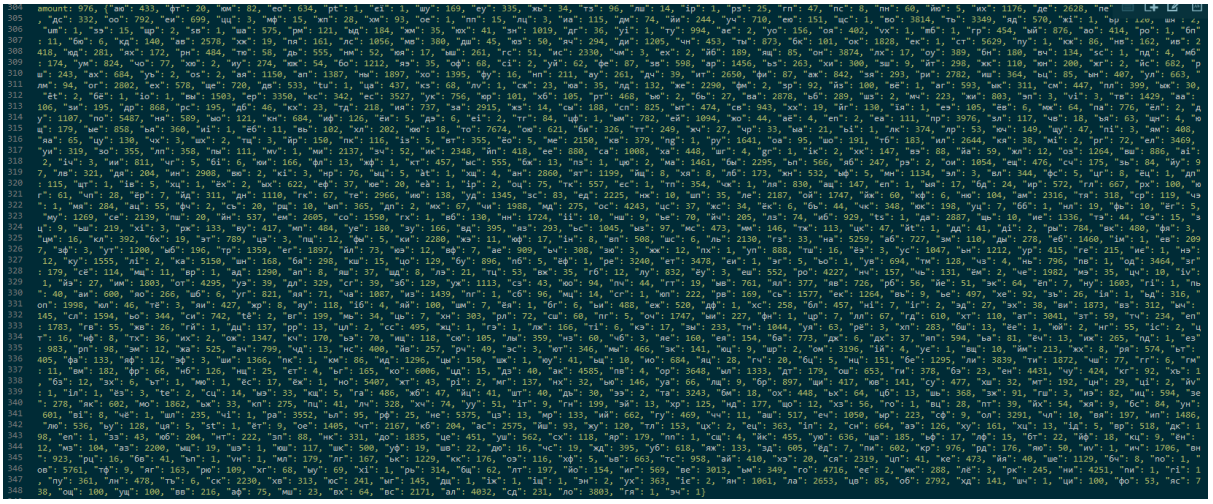
Частота: строка 66-300

H1 bigram1 = 8.287495344039405

H2 bigram1 = 4.143747672019702

1.3 Кількість та частота біграм з перетенами без пробілів:

Кількість: строка 304-348



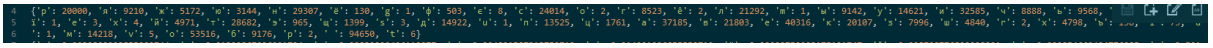
Частота: строка 349-602

H1 bigram2 = 14.578356699150458

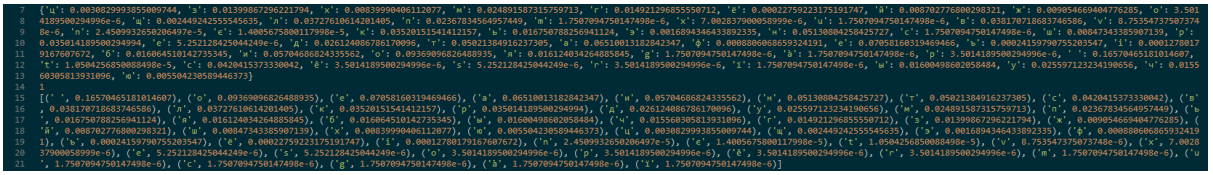
H2 bigram2 = 7.289178349575229

1.4 Кількість та частота літер, якщо вважати, що пробіл входить до алфавіту:

Кількість: строка 4-6



Частота: строка 7-21

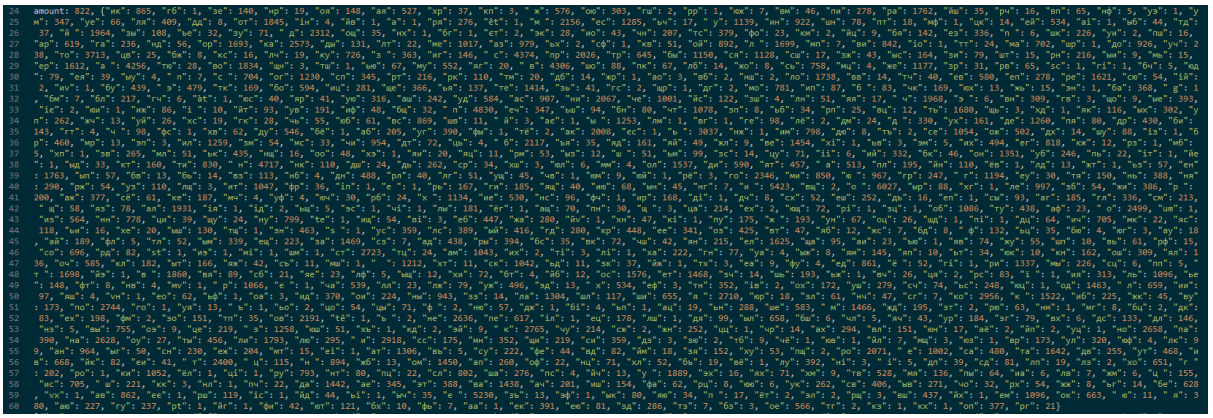


H1 chars = 4.385256604466399

H2 chars = 2.1926283022331994

1.5 Кількість та частота біграм з перетенами і без пробілів:

Кількість: строка 24-60



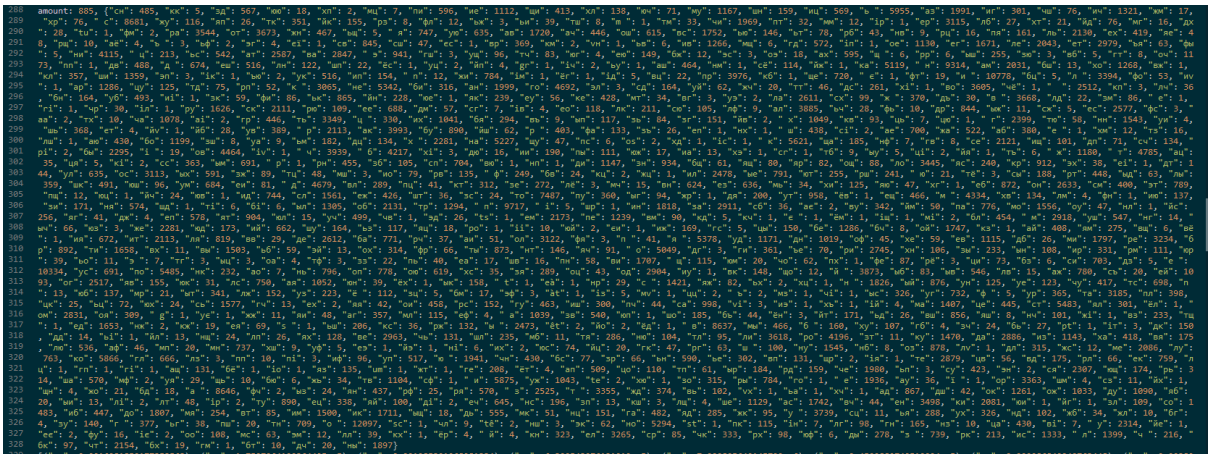
Частота: строка 61-284

H1 bigram1 = 7.947949394431538

H2 bigram1 = 3.973974697215769

1.6 Кількість та частота біграм з перетенами та з пробілами:

Кількість: строчка 288-328

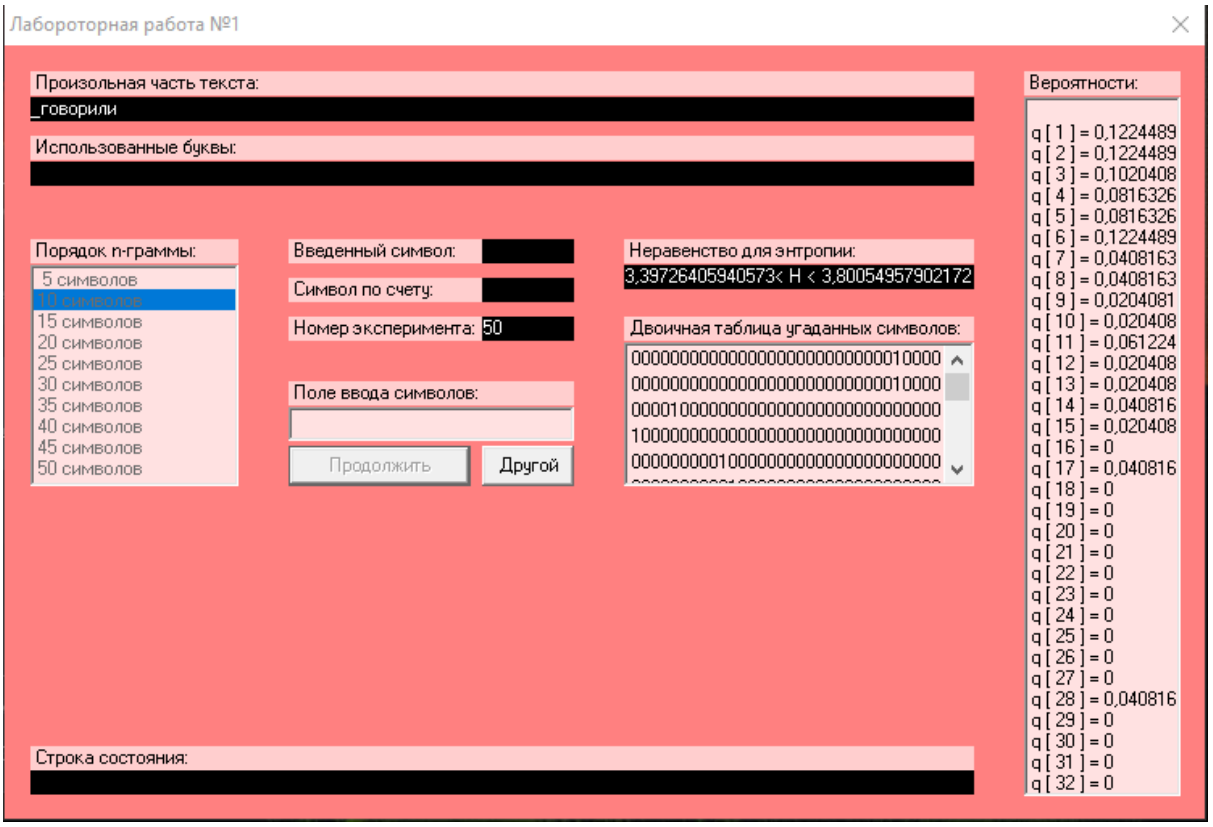


Частота: строчка 329-559

H1 bigram2 = 13.899736404213 69

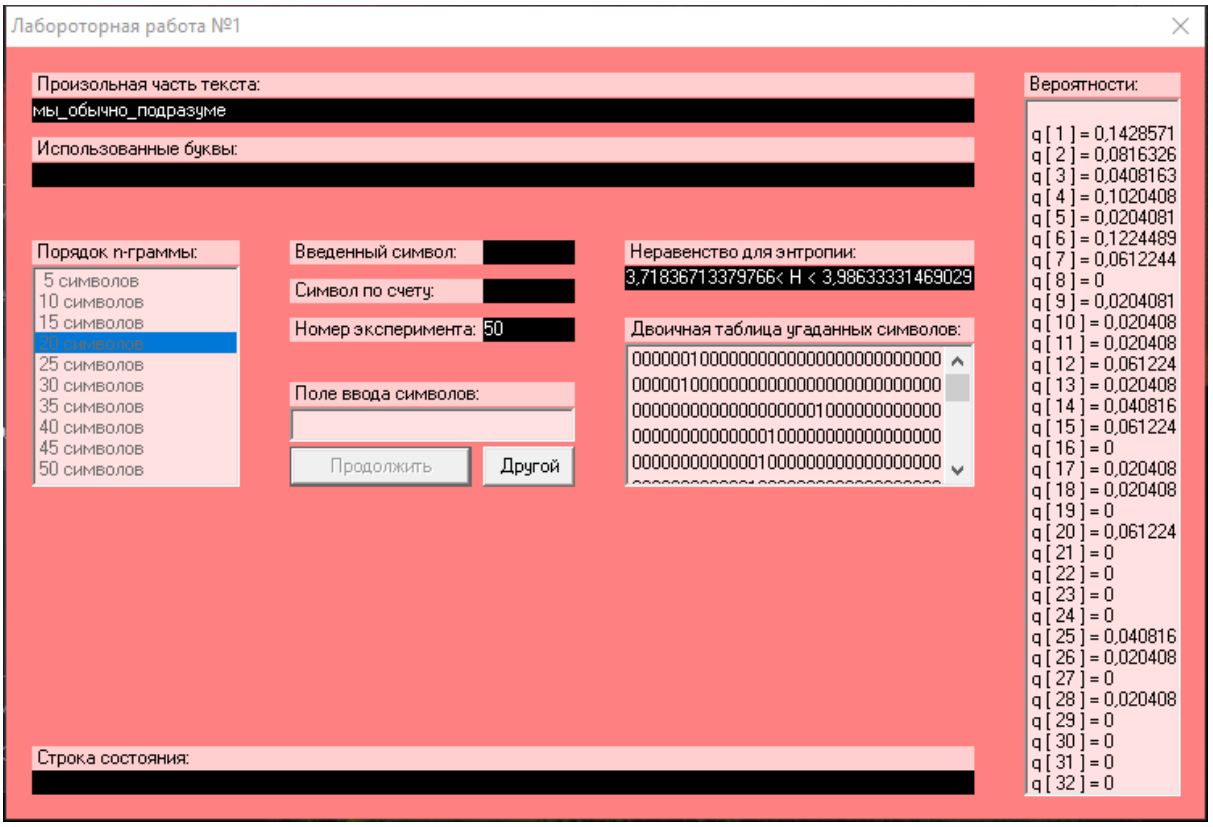
H2 bigram2 = 6.949868202106845

2. H(10)



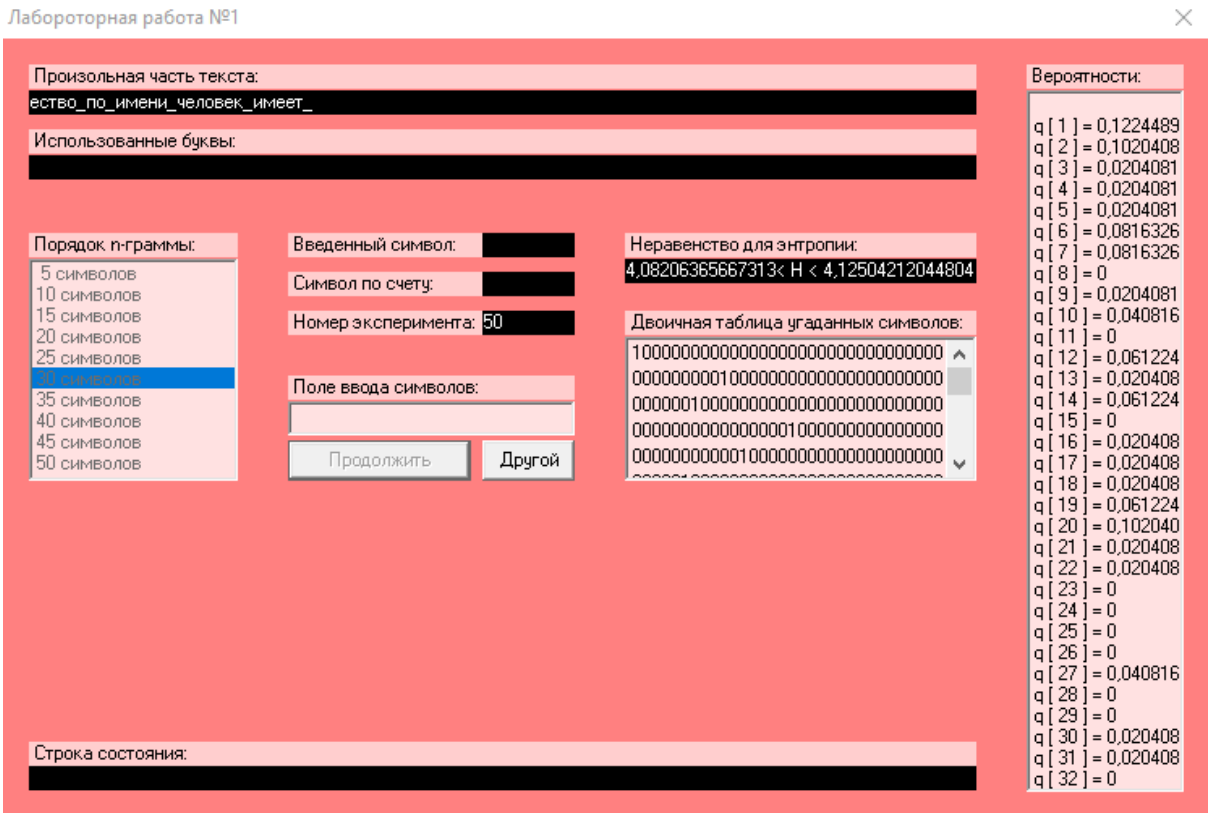
R = 0,40328551961599

H(20)



R = 0,26796618089263

H(30)



R = 0,04297846377491

Висновок:

Виявилось, що у творі “Мертві душі” : найбільш популярний символ це розділовий знак, воно в принципі зрозуміло чому. Другий за популярністю йде символ “о”, вражає що за 181 рік популярність символу в мові не змінилася.