Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
ФБ-21 Редько-Шпак Р.А.
ФБ-21 Серяков В.Л.

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.

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

Пункт №1

Для виконання лабораторної роботи було розроблено програму на мові Python, яка виконує аналіз тексту та обчислює різні статистичні показники. Основні етапи роботи коду:

- ✧ Зчитування тексту з файлу (в нашому випадку text.txt - Майстер та Маргарита, бо за умовою необхідно було проаналізувати текст >1 мб)
- ✧ Попередня обробка тексту (видалення зайвих символів, приведення до нижнього регістру)
- ✧ Створення версії тексту без пробілів (в подальшому вона буде збережена як text_no_spaces.txt)
- ✧ Підрахунок частот символів та біграм (для текстів з пробілами та без)
- ✧ Обчислення ентропії H1 та H2
- ✧ Обчислення надлишковості
- ✧ Візуалізація результатів у вигляді матриць біграм (хітмапами)
- ✧ Збереження результатів у файли (можливо, для подальшої обробки та аналізу)

Аналіз (текст з пробілами)

Консольний вивід (скорочено)

```
===== АНАЛІЗ ТЕКСТУ З ПРОБІЛАМИ =====
|] Загальна кількість символів: 712213
||] Загальна кількість біграм: 712212

[x] Частоти букв [x]:
 : 0.1602441966  о: 0.0929160237  а: 0.0725400969  е: 0.0682717109  и: 0.0574729751
н: 0.0538490592  т: 0.0506365371  л: 0.0441230362  с: 0.0425687259  р: 0.0398588624
в: 0.0395471579  к: 0.0307155303  у: 0.0252382363  м: 0.0252017304  п: 0.0238748801
д: 0.0236052979  г: 0.0160836716  я: 0.0160134679  ь: 0.0151906803  з: 0.0149281184
ы: 0.0145251491  ч: 0.0133541511  б: 0.0130045366  й: 0.0098439652  ж: 0.0075665566
ш: 0.0073320762  х: 0.0069094498  ю: 0.0044340668  щ: 0.0029710213  ц: 0.0027856835
э: 0.0025792846  ф: 0.0018140641

[x/x] Біграми (перетинаючі) [x/x]:
о : 0.0206595789  а : 0.0175046194  п: 0.0166144350  в: 0.0160022578  и : 0.0159713681
... ... ...
фб: 0.0000014041  юя: 0.0000014041  тэ: 0.0000014041  цб: 0.0000014041  фп: 0.0000014041

[x|x] Біграми (не перетинаючі) [x|x]:
о : 0.0207775213  а : 0.0173908892  п: 0.0167450141  е : 0.0160317434  и : 0.0158744868
... ... ...
лэ: 0.0000028082  юя: 0.0000028082  рф: 0.0000028082  йа: 0.0000028082  тэ: 0.0000028082

[SAVE] Збереження даних [SAVE]

В папку "output\з_пробілами" збережено:
+] Файл частоти_букв_з_пробілами.xlsx
+] Файл частоти_біграм_перетинаючі_з_пробілами.xlsx
+] Файл частоти_біграм_не_перетинаючі_з_пробілами.xlsx
+] Файл сортовані_біграми_перетинаючі_з_пробілами.xlsx
+] Файл сортовані_біграми_не_перетинаючі_з_пробілами.xlsx

В папку "output\біграм_матриці" збережено:
+] Файл матриця_біграми_перетинаючі_для_тексту_з_пробілами.png
+] Файл матриця_біграми_не_перетинаючі_для_тексту_з_пробілами.png
```

- **Знайдено загальну к-сть символів та біграм:**

```
|] Загальна кількість символів: 712213
||] Загальна кількість біграм: 712212
```

- **Частоти букв (а також повні значення збереження в excel файл):**

```
[x] Частоти букв [x]:
 : 0.1602441966  о: 0.0929160237  а: 0.0725400969  е: 0.0682717109  и: 0.0574729751
н: 0.0538490592  т: 0.0506365371  л: 0.0441230362  с: 0.0425687259  р: 0.0398588624
в: 0.0395471579  к: 0.0307155303  у: 0.0252382363  м: 0.0252017304  п: 0.0238748801
д: 0.0236052979  г: 0.0160836716  я: 0.0160134679  ь: 0.0151906803  з: 0.0149281184
ы: 0.0145251491  ч: 0.0133541511  б: 0.0130045366  й: 0.0098439652  ж: 0.0075665566
ш: 0.0073320762  х: 0.0069094498  ю: 0.0044340668  щ: 0.0029710213  ц: 0.0027856835
э: 0.0025792846  ф: 0.0018140641
```

- **Частоти біграм (повні значення збереження в excel файл):**

/ Перетинаючі (перші 5 та останні 5 за частотою):

```
[x/x] Біграми (перетинаючі) [x/x]:
о : 0.0206595789  а : 0.0175046194  п: 0.0166144350  в: 0.0160022578  и : 0.0159713681
... ... ...
фб: 0.0000014041  юя: 0.0000014041  тэ: 0.0000014041  цб: 0.0000014041  фп: 0.0000014041
```

| Не перетинаючі (перші 5 та останні 5 за частотою):

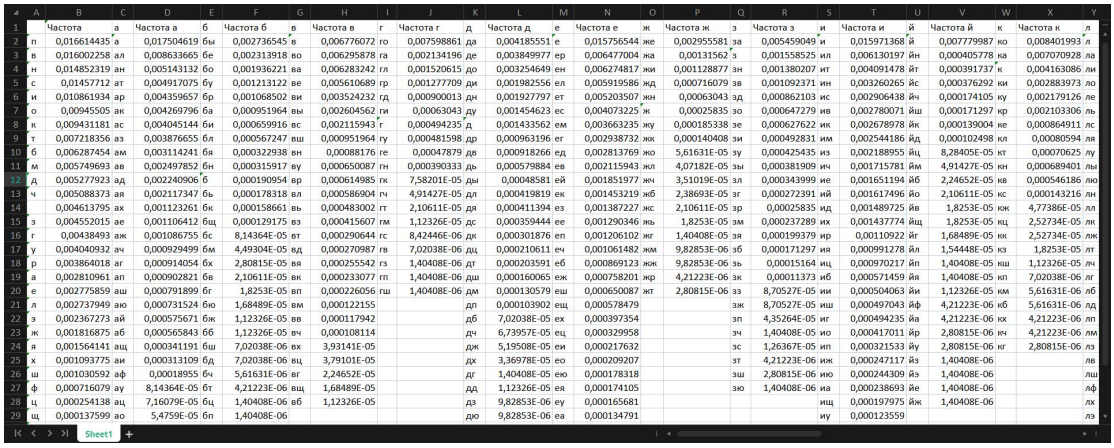
```
[x|x] Біграми (не перетинаючі) [x|x]:
о : 0.0207775213  а : 0.0173908892  п: 0.0167450141  е : 0.0160317434  и : 0.0158744868
... ... ...
лэ: 0.0000028082  юя: 0.0000028082  рф: 0.0000028082  йа: 0.0000028082  тэ: 0.0000028082
```

- **Таблиці з сортованими біграмами за першими буквами**

Зі вже отриманих мною даних біграм було створено ексель таблиці (і для перетинаючих, і для не перетинаючих). Їх структура така - перший рядок колонки - це буква алфавіту (а, б, в, ... я) або (, а, б, в, ... , я - у випадку тексту з пробілами), а далі вниз ідуть всі біграми, які починаються з цієї літери. І, що важливо, вони відсортовані за частотою (тобто від найпоширеніших до рідких), а також для зручності в сусідньому стовпичку пишеться частота, що відповідає цій біграми.

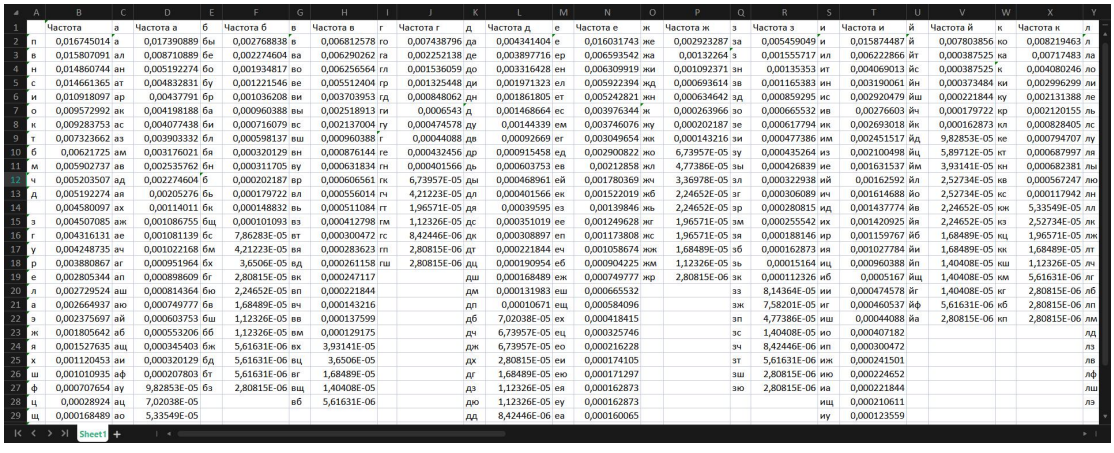
Це було створено для того, щоб було наочніше зрозуміло і можна було самостійно подивтися от для якої букви які біграми найпопулярніші, як навпаки рідкі, бо в сирому вигляді - дуже важко розібратися.

Сортовані біграми перетинаючі з пробілами



Частота	а	Частота а	б	Частота б	в	Частота в	г	Частота г	д	Частота д	е	Частота е	ж	Частота ж	з	Частота з	и	Частота и	й	Частота й	к	Частота к	л
п	0,016614435	а	0,017504619	бы	0,002736545	в	0,006776072	го	0,007598861	да	0,004185551	е	0,015756544	же	0,002955581	за	0,005459049	и	0,015971368	й	0,007779987	ко	0,008401993	л
в	0,016002258	ал	0,008633665	бе	0,002313918	во	0,006295878	га	0,002134196	де	0,003849977	ер	0,006477004	жа	0,00131562	з	0,001558525	ил	0,006130197	йн	0,000405778	ка	0,007070928	ла
н	0,014852319	ан	0,005143132	бо	0,001936221	ва	0,006283242	гл	0,001520615	до	0,003254649	ен	0,006274817	жи	0,001128877	зн	0,001380207	ит	0,004091478	йт	0,000391737	к	0,004163086	ли
с	0,01457712	ат	0,004917075	бу	0,001213122	ве	0,005610689	гр	0,001277709	ди	0,001982556	ел	0,005919586	жд	0,000716079	зв	0,001092371	ин	0,003260265	йс	0,000376292	ки	0,002883973	ло
и	0,010861934	ар	0,004359657	бр	0,001068502	ви	0,003524232	гд	0,000900013	дн	0,001927797	ет	0,005203507	жн	0,00063043	зд	0,000862103	ис	0,002906438	йч	0,000174105	ку	0,002179126	ле
о	0,00945505	ак	0,004269796	ба	0,000951964	вы	0,002604562	ги	0,00063043	ду	0,001454623	ес	0,004073225	ж	0,00025835	зо	0,000647279	ив	0,002780071	йш	0,000171297	кр	0,002103306	ль
к	0,009431181	ас	0,004045144	би	0,000659916	вс	0,002115943	г	0,000494235	д	0,001433562	ем	0,003663235	жу	0,000185338	зе	0,000627622	ик	0,002678978	йк	0,000139004	ке	0,000864911	лс
т	0,007218356	аз	0,003876655	бл	0,000567247	вш	0,000951964	гу	0,000481598	др	0,000963196	ег	0,002938732	жж	0,000140408	зи	0,000492831	им	0,002544186	йд	0,000102498	кл	0,00080594	ля
б	0,006287454	ам	0,003114241	бя	0,000322938	вн	0,00088176	ге	0,00047879	дв	0,000918266	ед	0,002813769	жо	5,61631E-05	зу	0,000425435	из	0,002188955	йц	8,28405E-05	кт	0,00070625	лу
м	0,005749693	ав	0,002497852	бн	0,000315917	ву	0,000650087	гн	0,000390333	дь	0,000579884	ев	0,002115943	жл	4,07182E-05	зы	0,000381909	ич	0,001715781	йм	4,91427E-05	кн	0,000689401	лы
д	0,005277923	ад	0,002240906	б	0,000190954	вр	0,000614985	гк	7,58201E-05	ды	0,00048581	ей	0,001851977	жч	3,51019E-05	зл	0,000343999	ие	0,001651194	йб	2,24652E-05	кв	0,000546186	лю
ч	0,005088373	ая	0,002117347	бь	0,000178318	вл	0,000586904	гч	4,91427E-05	дл	0,000419819	ек	0,001453219	жб	2,38693E-05	зг	0,000272391	ий	0,001617496	йо	2,10611E-05	кс	0,000143216	лн
	0,004613795	ах	0,001123261	бк	0,000158661	вь	0,000483002	гт	2,10611E-05	дя	0,000411394	ез	0,001387227	жс	2,10611E-05	зр	0,00025835	ид	0,001489725	йв	1,8253E-05	кю	4,77386E-05	лл
з	0,004552015	ае	0,001106412	бщ	0,000129175	вз	0,000415607	гм	1,12326E-05	дс	0,000359444	ее	0,001290346	жь	1,8253E-05	зм	0,000237289	их	0,001437774	йщ	1,8253E-05	кц	2,52734E-05	лк
г	0,00438493	аж	0,001086755	бс	8,14364E-05	вт	0,000290644	гс	8,42446E-06	дк	0,000301876	еп	0,001206102	жг	1,40408E-05	зя	0,000199379	ир	0,0010922	йг	1,68489E-05	кк	2,52734E-05	лж
у	0,004040932	ач	0,000929499	бм	4,49304E-05	вд	0,000270987	гз	7,02038E-06	дц	0,000210611	еч	0,001061482	жм	9,82853E-06	зб	0,000171297	ия	0,000991278	йл	1,54448E-05	кз	1,8253E-05	лт
р	0,003864018	аг	0,000914054	бх	2,80815E-05	вя	0,000255542	гз	1,40408E-06	дт	0,000203591	еб	0,000869123	жж	9,82853E-06	зь	0,00015164	иц	0,000970217	йп	1,40408E-05	кш	1,12326E-05	лч
а	0,002810961	ап	0,000902821	бв	2,10611E-05	вк	0,000233077	гп	1,40408E-06	дш	0,000160065	еж	0,000758201	жр	4,21223E-06	зк	0,00011373	иб	0,000571459	йя	1,40408E-05	кп	7,02038E-06	лг
е	0,002775859	аш	0,000791899	бг	1,8253E-05	вп	0,000226056	гш	1,40408E-06	дм	0,000130579	еш	0,000650087	жт	2,80815E-06	зз	8,70527E-05	ии	0,000504063	йн	1,12326E-05	км	5,61631E-06	лб
л	0,002737949	аю	0,000731524	бю	1,68489E-05	вм	0,000122155			дп	0,000103902	ещ	0,000578479			зж	8,70527E-05	иш	0,000497043	йф	4,21223E-06	кб	5,61631E-06	лд
э	0,002367273	ай	0,000575671	бж	1,12326E-05	вв	0,000117942			дб	7,02038E-05	ех	0,000397354			зп	4,35264E-05	иг	0,000494235	йа	4,21223E-06	кж	4,21223E-06	лп
ж	0,001816875	аб	0,000565843	бб	1,12326E-05	вч	0,000108114			дч	6,73957E-05	ец	0,000329958			зч	1,40408E-05	ио	0,000417011	йр	2,80815E-06	нч	4,21223E-06	лм
я	0,001564141	ащ	0,000341191	бш	7,02038E-06	вх	3,93141E-05			дж	5,19508E-05	еи	0,000217632			зс	1,26367E-05	ип	0,000321533	йу	2,80815E-06	нг	2,80815E-06	лз
х	0,001093775	аи	0,000313109	бд	7,02038E-06	вц	3,79101E-05			дх	3,36978E-05	ео	0,000209207			зт	4,21223E-06	иж	0,000247117	йз	1,40408E-06			лв
ш	0,001030592	аф	0,000189555	бч	5,61631E-06	вг	2,24652E-05			дг	1,40408E-05	ею	0,000178318			зш	2,80815E-06	ию	0,000244309	йэ	1,40408E-06			лш
ф	0,000716079	ау	8,14364E-05	бт	4,21223E-06	вщ	1,68489E-05			дд	1,12326E-05	ея	0,000174105			зю	1,40408E-06	иа	0,000238693	йе	1,40408E-06			лф
ц	0,000254138	ац	7,16079E-05	бц	1,40408E-06	вб	1,12326E-05			дз	9,82853E-06	еу	0,000165681					ищ	0,000197975	йж	1,40408E-06			лх
щ	0,000137599	ао	5,4759E-05	бп	1,40408E-06					дю	9,82853E-06	еа	0,000134791					иу	0,00012355					лэ

Сортовані біграми не перетинаючі з пробілами



Частота	а	Частота а	б	Частота б	в	Частота в	г	Частота г	д	Частота д	е	Частота е	ж	Частота ж	з	Частота з	и	Частота и	й	Частота й	к	Частота к	л
п	0,016745014	а	0,017390889	бы	0,002768838	в	0,006812578	го	0,007438796	да	0,004341404	е	0,016031743	же	0,002923287	за	0,005459049	и	0,015874487	й	0,007803856	ко	0,008219463	л
в	0,015807091	ал	0,008710889	бе	0,002274604	ва	0,006290262	га	0,002252138	де	0,003897716	ер	0,006593542	жа	0,00132264	з	0,001555717	ил	0,006222866	йт	0,000387525	ка	0,00717483	ла
н	0,014860744	ан	0,005192274	бо	0,001934817	во	0,006256564	гл	0,001536059	до	0,003316428	ен	0,006309919	жи	0,001092371	зн	0,00135353	ит	0,004069013	йс	0,000387525	к	0,004080246	ло
с	0,014661365	ат	0,004832831	бу	0,001221546	ве	0,005512404	гр	0,001325448	ди	0,001971323	ел	0,005922394	жд	0,000693614	зв	0,001165383	ин	0,003190061	йн	0,000373484	ки	0,002996299	ли
и	0,010918097	ар	0,00437791	бр	0,001036208	ви	0,003703953	гд	0,000848062	дн	0,001861805	ет	0,005242821	жн	0,000634642	зд	0,000859295	ис	0,002920479	йч	0,000221844	ку	0,002131388	ле
о	0,009572992	ак	0,004198188	ба	0,000960388	вы	0,002518913	ги	0,0006543	д	0,001468664	ес	0,003976344	ж	0,000263966	зо	0,000665532	ив	0,00276603	йч	0,000179722	кр	0,002120155	ль
к	0,009283753	ас	0,004077438	би	0,000716079	вс	0,002137004	гу	0,000474578	ду	0,00144339	ем	0,003746076	жу	0,000202187	зе	0,000617794	ик	0,002693018	йк	0,000162873	кл	0,000828405	лс
т	0,007323662	аз	0,003903332	бл	0,000598137	вш	0,000960388	г	0,00044088	дв	0,00092669	ег	0,003049654	жж	0,000143216	зи	0,000477386	им	0,002451517	йд	9,82853E-05	ке	0,000794707	лу
б	0,00621725	ам	0,003176021	бя	0,000320129	вн	0,000876144	ге	0,000432456	дв	0,000915458	ед	0,002900822	жо	6,73957E-05	зу	0,000435264	из	0,002100498	йц	5,89712E-05	кт	0,000687997	ля
м	0,005902737	ав	0,002535762	бн	0,000311705	ву	0,000631834	гн	0,000401566	дь	0,000603753	ев	0,00212858	жл	4,77386E-05	зы	0,000426839	ие	0,001631537	йм	3,93141E-05	кн	0,000682381	лы
ч	0,005203507	ад	0,002274604	б	0,000202187	вр	0,000606561	гк	6,73957E-05	ды	0,000468961	ей	0,001780369	жч	3,36978E-05	зл	0,000322938	ий	0,00162592	йл	2,52734E-05	кв	0,000567247	лю
д	0,005192274	ая	0,00205276	бь	0,000179722	вл	0,000556014	гч	4,21223E-05	дл	0,000401566	ек	0,001522019	жб	2,24652E-05	зг	0,000306089	ич	0,001614688	йо	2,52734E-05	кс	0,000117942	лн
	0,004580097	ах	0,00114011	бк	0,000148832	вь	0,000511084	гт	1,96571E-05	дя	0,00039595	ез	0,00139846	жь	2,24652E-05	зр	0,000280815	ид	0,001437774	йв	2,24652E-05	кю	5,33549E-05	лл
з	0,004507085	аж	0,001086755	бщ	0,000101093	вз	0,000412798	гм	1,12326E-05	дс	0,000351019	ее	0,001249628	жг	1,96571E-05	зм	0,000255542	их	0,001420925	йг	2,24652E-05	кц	2,52734E-05	лк
г	0,004316131	ае	0,001081139	бс	7,86283E-05	вт	0,000300472	гс	8,42446E-06	дк	0,000308897	еп	0,001173808	жс	1,96571E-05	зя	0,000188146	ир	0,001159767	йб	1,68489E-05	кщ	1,96571E-05	лж
у	0,004248735	ач	0,001022168	бм	4,21223E-05	вя	0,000283623	гп	2,80815E-06	дт	0,000221844	еч	0,001058674	жн	1,68489E-05	зб	0,000162873	ия	0,001027784	йн	1,68489E-05	кч	1,68489E-05	лт
р	0,003880867	аг	0,000951964	бх	3,6506E-05	вя	0,000261158	гш	2,80815E-06	дц	0,000190954	еб	0,000904225	жм	1,12326E-05	зь	0,00015164	иц	0,000960388	йп	1,40408E-05	кш	1,12326E-05	лч
е	0,002805344	ап	0,000898609	бг	2,80815E-05	вк	0,000247117			дш	0,000168489	еж	0,000749777	жр	2,80815E-06	зк	0,000112326	иб	0,0005167	йщ	1,40408E-05	км	5,61631E-06	лг
л	0,002729524	аш	0,000814364	бю	2,24652E-05	вп	0,000221844			дм	0,000131983	еш	0,000665532			зз	8,14364E-05	ии	0,000474578	йг	1,40408E-05	кг	2,80815E-06	лд
а	0,002664937	аю	0,000749777	бв	1,68489E-05	вч	0,000143216			дп	0,00010671	ещ	0,000584096			зж	7,58201E-05	иг	0,000460537	йф	5,61631E-06	кб	2,80815E-06	лп
э	0,002375697	ай	0,000603753	бш	1,12326E-05	вв	0,000137599			дб	7,02038E-05	ех	0,000418415			зп	4,77386E-05	иш	0,00044088	йа	2,80815E-06	кп	2,80815E-06	лм
ж	0,001805642	аб	0,000553206	бб	1,12326E-05	вм	0,000129175			дч	6,73957E-05	ец	0,000325746			зс	1,40408E-05	ио	0,000407182					лд
я	0,001527635	ащ	0,000345403	бж	5,61631E-06	вх	3,93141E-05			дж	6,73957E-05	ео	0,000216228			зч	8,42446E-06	ип	0,000300472					лз
х	0,001120453	аи	0,000320129	бд	5,61631E-06	вц	3,6506E-05			дх	2,80815E-05	еи	0,000174105			зт	5,61631E-06	иж	0,000241501					лв
ш	0,001010935	аф	0,000207803	бт	5,61631E-06	вг	1,68489E-05			дг	1,68489E-05	ею	0,000171297			зш	2,80815E-06	ию	0,000224652					лф
ф	0,000707654	ау	9,82853E-05	бз	2,80815E-06	вщ	1,40408E-05			дз	1,12326E-05	ея	0,000162873			зю	2,80815E-06	иа	0,000221844					лш
ц	0,00028924	ац	7,02038E-05			вб	5,61631E-06			дд	1,12326E-05	еу	0,000162873					ищ	0,000210611					лх
щ	0,00016849	ао	5,3549E-05							дю	8,42446E-06	еа	0,000160065					иу	0,000123559					лэ

- **Логування про збереження результатів в excel:**

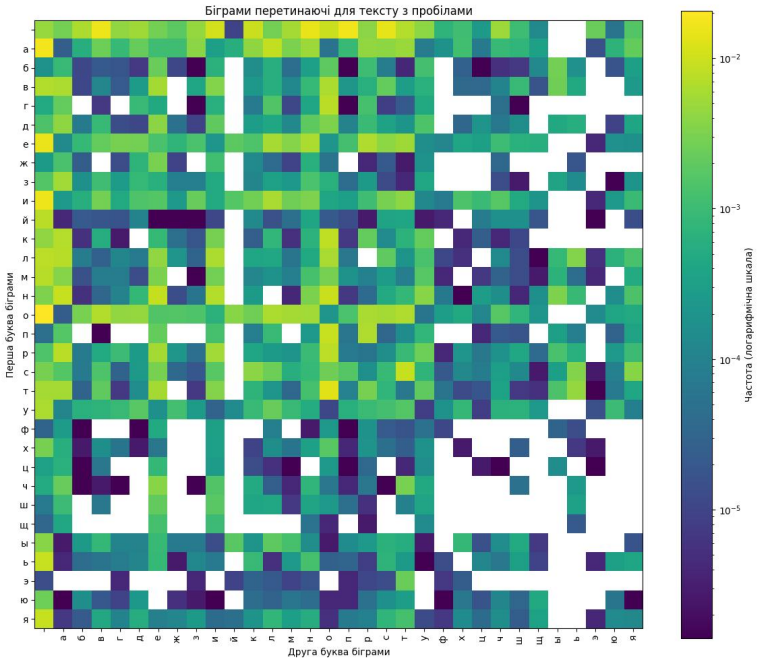
```
[SAVE] Збереження даних [SAVE]

В папку "output\з_пробілами" збережено:
+] Файл частоти_букв_з_пробілами.xlsx
+] Файл частоти_біграм_перетинаючі_з_пробілами.xlsx
+] Файл частоти_біграм_не_перетинаючі_з_пробілами.xlsx
+] Файл сортовані_біграми_перетинаючі_з_пробілами.xlsx
+] Файл сортовані_біграми_не_перетинаючі_з_пробілами.xlsx
```

➢ Матриці біграм

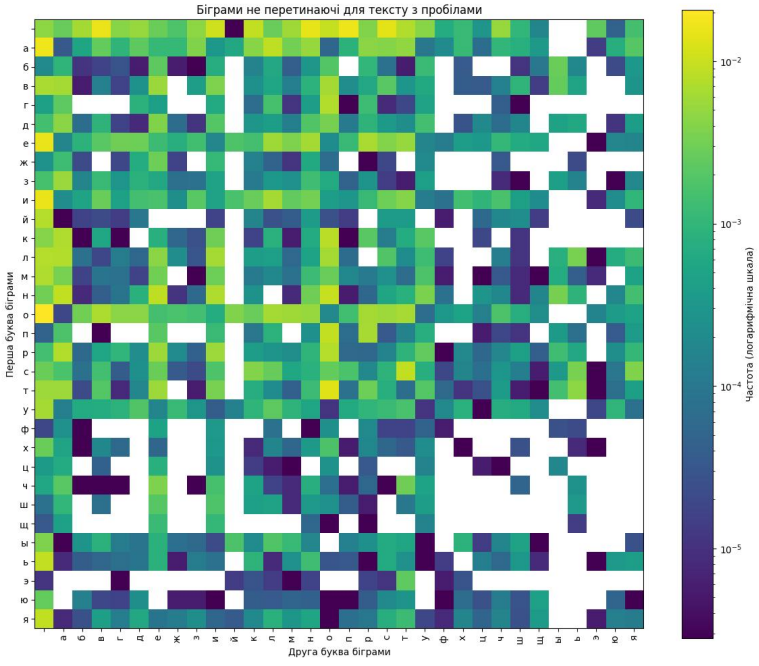
```
В папку "output\біграм_матриці" збережено:
+] Файл матриця_біграми_перетинаючі_для_тексту_з_пробілами.png
+] Файл матриця_біграми_не_перетинаючі_для_тексту_з_пробілами.png
```

В методичних вказівках було сказано, що таблицю частот біграм зручно подавати у вигляді квадратної матриці, індексованої першою та другою літерами біграм. Тому це було зроблено за допомогою хітмапи бібліотеки матплотліб:

Матриця біграми перетинаючі для тексту з пробілами



Матриця біграми не перетинаючі для тексту з пробілами



Аналіз (текст без пробілів)

За умовою завдання, треба було також проаналізувати текст без пробілів, тому були створені і використані відповідні ф-ії. Спочатку текст фільтрується за допомогою ф-ії filter_text() - переводить весь текст у нижній регістр, замінює 'ё' на 'е' та 'ъ' на 'ь', видаляє всі символи, крім букв та пробілів (за умовою). Вже після цього ф-ія remove_spaces() видаляє пробіли з тексту та ф-ія save_text_no_spaces_to_txt() - зберігає текст без пробілів у файл.

Консольний вивід (скорочено)

```
===== АНАЛІЗ ТЕКСТУ БЕЗ ПРОБІЛІВ =====
|] Загальна кількість символів: 598085
||] Загальна кількість біграм: 598084

[x] Частоти букв [x]:
о: 0.1106464800  а: 0.0863823704  е: 0.0812994808  и: 0.0684401047  н: 0.0641246646
т: 0.0602991214  л: 0.0525426988  с: 0.0506917913  р: 0.0474648252  в: 0.0470936405
к: 0.0365767408  у: 0.0300542565  м: 0.0300107844  п: 0.0284307414  д: 0.0281097168
г: 0.0191527960  я: 0.0190691959  ь: 0.0180894020  з: 0.0177767374  ы: 0.0172968725
ч: 0.0159024219  б: 0.0154860931  й: 0.0117224140  ж: 0.0090104249  ш: 0.0087312004
х: 0.0082279275  ю: 0.0052801859  щ: 0.0035379587  ц: 0.0033172542  э: 0.0030714698
ф: 0.0021602281

[x/x] Біграми (перетинаючі) [x/x]:
то: 0.0162251456  но: 0.0118227540  ст: 0.0111823757  на: 0.0109683590  по: 0.0109048227
... ... ...
фя: 0.0000016720  фб: 0.0000016720  пд: 0.0000016720  эб: 0.0000016720  шч: 0.0000016720

[x|x] Біграми (не перетинаючі) [x|x]:
то: 0.0164190983  но: 0.0116271293  ст: 0.0111589676  по: 0.0110853994  на: 0.0110553033
... ... ...
фя: 0.0000033440  фб: 0.0000033440  дщ: 0.0000033440  шч: 0.0000033440  эп: 0.0000033440

[SAVE] Збереження даних [SAVE]

В папку "output\без_пробілів" збережено:
+] Файл частоти_букв_без_пробілів.xlsx
+] Файл частоти_біграм_перетинаючі_без_пробілів.xlsx
+] Файл частоти_біграм_не_перетинаючі_без_пробілів.xlsx
+] Файл сортовані_біграми_перетинаючі_без_пробілів.xlsx
+] Файл сортовані_біграми_не_перетинаючі_без_пробілів.xlsx

В папку "output\біграм_матриці" збережено:
+] Файл матриця_біграми_перетинаючі_для_тексту_без_пробілів.png
+] Файл матриця_біграми_не_перетинаючі_для_тексту_без_пробілів.png
```

- **Знайдено загальну к-сть символів та біграм:**

```
|] Загальна кількість символів: 598085
||] Загальна кількість біграм: 598084
```

Як видно, кількість букв та біграм зменшилася більше, ніж на 100 тисяч, після обробки та видалення пробілів.

- **Частоти букв (повні значення збереження в excel файл):**

```
[x] Частоти букв [x]:
о: 0.1106464800  а: 0.0863823704  е: 0.0812994808  и: 0.0684401047  н: 0.0641246646
т: 0.0602991214  л: 0.0525426988  с: 0.0506917913  р: 0.0474648252  в: 0.0470936405
к: 0.0365767408  у: 0.0300542565  м: 0.0300107844  п: 0.0284307414  д: 0.0281097168
г: 0.0191527960  я: 0.0190691959  ь: 0.0180894020  з: 0.0177767374  ы: 0.0172968725
ч: 0.0159024219  б: 0.0154860931  й: 0.0117224140  ж: 0.0090104249  ш: 0.0087312004
х: 0.0082279275  ю: 0.0052801859  щ: 0.0035379587  ц: 0.0033172542  э: 0.0030714698
ф: 0.0021602281
```

- **Частоти біграм (а також повні значення збереження в excel файл):**

/ Перетинаючі (перші 5 та останні 5 за частотою):

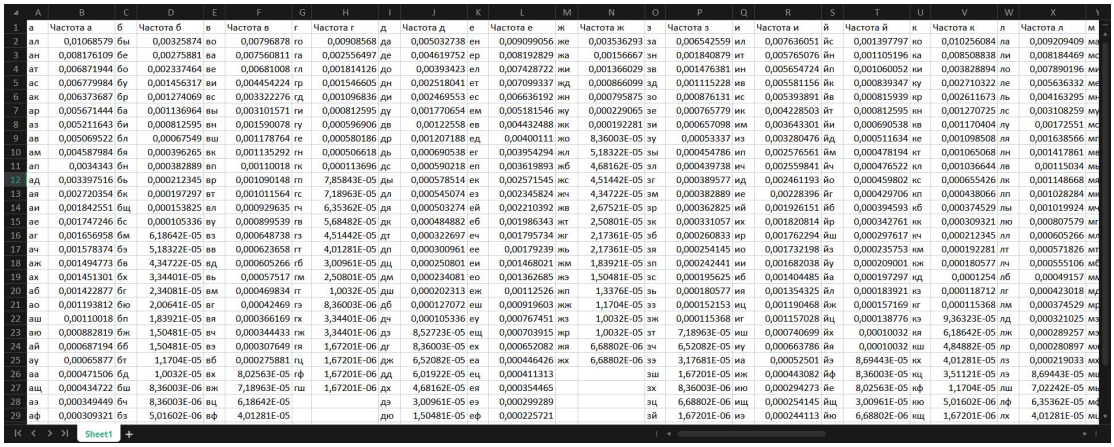
```
[х/х] Біграми (перетинаючі) [х/х]:
то: 0.0162251456  но: 0.0118227540  ст: 0.0111823757  на: 0.0109683590  по: 0.0109048227
... ... ...
фя: 0.0000016720  фб: 0.0000016720  пд: 0.0000016720  эб: 0.0000016720  шч: 0.0000016720
```

| Не перетинаючі (перші 5 та останні 5 за частотою):

```
[х|х] Біграми (не перетинаючі) [х|х]:
то: 0.0164190983  но: 0.0116271293  ст: 0.0111589676  по: 0.0110853994  на: 0.0110553033
... ... ...
фя: 0.0000033440  фб: 0.0000033440  дщ: 0.0000033440  шч: 0.0000033440  эп: 0.0000033440
```

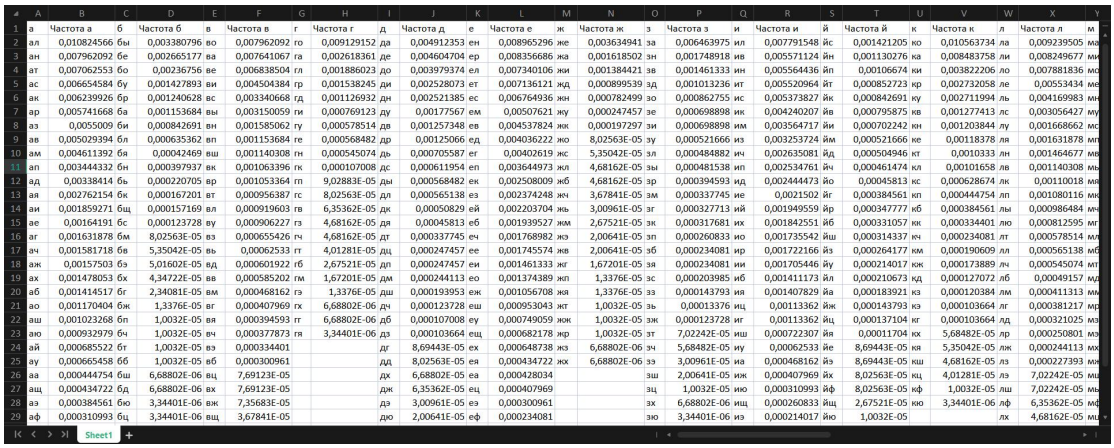
- **Таблиці з сортованими біграмами за першими буквами**

Сортовані біграми перетинаючі без пробілів (



	А	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	а	Частота а	б	Частота б	в	Частота в	г	Частота г	д	Частота д	е	Частота е	ж	Частота ж	з	Частота з	и	Частота и	й	Частота й	к	Частота к	л	Частота л	м
2	ал	0,01068579	бы	0,00325874	во	0,00796878	го	0,00908568	да	0,005032738	ен	0,009099056	же	0,003536293	за	0,006542559	ил	0,007636051	йс	0,001397797	ко	0,010256084	ла	0,009209409	ма
3	ан	0,008176109	бе	0,00275881	ва	0,007560811	га	0,002556497	де	0,004619752	ер	0,008192829	жа	0,00156667	зн	0,001840879	ит	0,005765076	йн	0,001105196	ка	0,008508838	ли	0,008184469	мс
4	ат	0,006871944	бо	0,002337464	ве	0,00681008	гл	0,001814126	до	0,00393423	ел	0,007428722	жи	0,001366029	зв	0,001476381	ин	0,005654724	йп	0,001060052	ки	0,003828894	ло	0,007890196	ми
5	ас	0,006779984	бу	0,001456317	ви	0,004454224	гр	0,001546605	дн	0,002518041	ет	0,007099337	жд	0,000866099	зд	0,001115228	ив	0,005581156	йк	0,000839347	ку	0,002710322	ле	0,005636332	ме
6	ак	0,006373687	бр	0,001274069	вс	0,003322276	гд	0,001096836	ди	0,002469553	ес	0,006636192	жн	0,000795875	зо	0,000876131	ис	0,005393891	йв	0,000815939	кр	0,002611673	ль	0,004163295	мн
7	ар	0,005671444	ба	0,001136964	вы	0,003101571	ги	0,000812595	ду	0,001770654	ем	0,005181546	жу	0,000229065	зе	0,000765779	ик	0,004228503	йт	0,000812595	кн	0,001270725	лс	0,003108259	му
8	аэ	0,005211643	би	0,000812595	вн	0,001590078	гу	0,000596906	дв	0,00122558	ев	0,004432488	жк	0,000192281	зи	0,000657098	им	0,003643301	йи	0,000690538	кв	0,001170404	лу	0,00172551	мс
9	ав	0,005069522	бл	0,00067549	вш	0,001178764	ге	0,000580186	др	0,001207188	ед	0,00400111	жо	8,36003E-05	зу	0,00053337	из	0,003280476	йд	0,000511634	ке	0,001098508	ля	0,001638566	мт
10	ам	0,004587984	ва	0,000396265	вк	0,001135292	гн	0,000506618	дь	0,000690538	ег	0,003954294	жл	5,18322E-05	зы	0,000454786	ип	0,002576561	йм	0,000478194	кт	0,001065068	лн	0,001417861	мф
11	ап	0,0034343	бн	0,000382889	вп	0,00110018	гк	0,000113696	дс	0,000590218	еп	0,003619893	жб	4,68162E-05	зл	0,000439738	ич	0,002559841	йч	0,000476522	кл	0,001036644	лв	0,00115034	мь
12	ад	0,003397516	бь	0,000212345	вр	0,001090148	гп	7,85843E-05	ды	0,000578514	ек	0,002571545	жс	4,51442E-05	зг	0,000389577	ид	0,002461193	йо	0,00045902	кс	0,000655426	лк	0,001148668	ма
13	ая	0,002720354	бк	0,000197297	вт	0,001011564	гс	7,18963E-05	дл	0,000545074	ез	0,002345824	жч	4,34722E-05	зм	0,000382889	ие	0,00228396	йг	0,000429706	кп	0,000438066	лп	0,001028284	мк
14	аи	0,001842551	бщ	0,000153825	вл	0,000929635	гч	6,35362E-05	дя	0,000503274	ей	0,002210392	жв	2,67521E-05	зр	0,000362825	ий	0,001926151	йб	0,000394593	кб	0,000374529	лы	0,001019924	мч
15	ае	0,001747246	бс	0,000105336	ву	0,000899539	гв	5,68482E-05	дк	0,000484882	еб	0,001986343	жт	2,50801E-05	зк	0,000331057	их	0,001820814	йр	0,000342761	кк	0,000309321	лю	0,000807579	мт
16	аг	0,001656958	бм	6,18642E-05	вз	0,000648738	гз	4,51442E-05	дт	0,000322697	еч	0,001795734	жг	2,17361E-05	зб	0,000260833	ир	0,001762294	йш	0,000297617	кч	0,000212345	лл	0,000605266	мл
17	ач	0,001578374	бз	5,18322E-05	вв	0,000623658	гт	4,01281E-05	дп	0,000300961	ее	0,00179239	жь	2,17361E-05	зя	0,000254145	ио	0,001732198	йз	0,000235753	км	0,000192281	лт	0,000571826	мт
18	аж	0,001494773	бя	4,34722E-05	вд	0,000605266	гб	3,00961E-05	дц	0,000250801	еи	0,001468021	жм	1,83921E-05	зп	0,000242441	ии	0,001682038	йу	0,00020901	юк	0,000180577	лч	0,000555106	мс
19	ах	0,001451301	бх	3,34401E-05	вь	0,00057517	гм	2,50801E-05	дм	0,000234081	ео	0,001362685	жз	1,50481E-05	зс	0,000195625	иб	0,001404485	йа	0,000197297	кд	0,0001254	лб	0,00049157	мж
20	аб	0,001422877	бг	2,34081E-05	вм	0,000469834	гг	1,0032E-05	дш	0,000202313	еж	0,00112526	жп	1,3376E-05	зь	0,000180577	ия	0,001354325	йл	0,000183921	кз	0,000118712	лг	0,000423018	мд
21	ао	0,001193812	бю	2,00641E-05	вг	0,00042469	гэ	8,36003E-06	дб	0,000127072	еш	0,000919603	жж	1,1704E-05	зз	0,000152153	иц	0,001190468	йж	0,000157169	кг	0,000115368	лм	0,000374529	мр
22	аш	0,00110018	бп	1,83921E-05	вя	0,000366169	гх	3,34401E-06	дч	0,000105336	еу	0,000767451	жэ	1,0032E-05	зж	0,000115368	иг	0,001157028	йц	0,000138776	кэ	9,36323E-05	лд	0,000321025	мэ
23	аю	0,000882819	бж	1,50481E-05	вч	0,000344433	гж	3,34401E-06	дз	8,52723E-05	ещ	0,000703915	жр	1,0032E-05	зт	7,18963E-05	иш	0,000740699	йх	0,00010032	кя	6,18642E-05	лж	0,000289257	мэ
24	ай	0,000687194	бб	1,50481E-05	вэ	0,000307649	гя	1,67201E-06	дг	8,36003E-05	ех	0,000652082	жя	6,68802E-06	зч	6,52082E-05	иу	0,000663786	йя	0,00010032	кш	4,84882E-05	лр	0,000280897	мю
25	ау	0,00065877	бт	1,1704E-05	вб	0,000275881	гц	1,67201E-06	дж	6,52082E-05	ea	0,000446426	жх	6,68802E-06	зэ	3,17681E-05	иа	0,00052501	йэ	8,69443E-05	кх	4,01281E-05	лз	0,000219033	мх
26	аа	0,000471506	бд	1,0032E-05	вх	8,02563E-05	гф	1,67201E-06	дд	6,01922E-05	ец	0,000411313			зш	1,67201E-05	иж	0,000443082	йф	8,36003E-05	кц	3,51121E-05	лэ	8,69443E-05	мш
27	ащ	0,000434722	бш	8,36003E-06	вж	7,18963E-05	гш	1,67201E-06	дх	4,68162E-05	ея	0,000354465			зх	8,36003E-06	ию	0,000294273	йе	8,02563E-05	кф	1,1704E-05	лш	7,02242E-05	мы
28	аэ	0,000349449	бч	8,36003E-06	вц	6,18642E-05			дэ	3,00961E-05	еэ	0,000299289			зц	6,68802E-06	ищ	0,000254145	йщ	3,00961E-05	кю	5,01602E-06	лф	6,35362E-05	мэ
29	аф	0,000309321	бз	5,01602E-06	вф	4,01281E-05			дю	1,50481E-05	еф	0,000225721			зй	1,67201E-06	из	0,000244113	йю	6,68802E-06	кщ	1,67201E-06	лх	4,01281E-05	мь

Сортовані біграми не перетинаючі без пробілів



	А	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	а	Частота а	б	Частота б	в	Частота в	г	Частота г	д	Частота д	е	Частота е	ж	Частота ж	з	Частота з	и	Частота и	й	Частота й	к	Частота к	л	Частота л	м
2	ал	0,010824566	бы	0,003380796	во	0,007962092	го	0,009129152	да	0,004912353	ен	0,008965296	же	0,003634941	за	0,006463975	ил	0,007791548	йс	0,001421205	ко	0,010563734	ла	0,009239505	ма
3	ан	0,007962092	бе	0,002665177	ва	0,007641067	га	0,002618361	де	0,004604704	ер	0,008356686	жа	0,001618502	зн	0,001748918	ив	0,005571124	йн	0,001130276	ка	0,008483758	ли	0,008249677	му
4	ат	0,007062553	бо	0,00236756	ве	0,006838504	гл	0,001886023	до	0,003979374	ел	0,007340106	жи	0,001384421	зв	0,001461333	ин	0,005564436	йп	0,00106674	ки	0,003822206	ло	0,007881836	ми
5	ас	0,006654584	бу	0,001427893	ви	0,004504384	гр	0,001538245	дн	0,002528073	ет	0,007136121	жд	0,000899539	зд	0,001013236	ив	0,005520964	йт	0,000852723	кр	0,002732058	ле	0,00553434	ме
6	ак	0,006239926	бр	0,001240628	вс	0,003340668	гд	0,001126932	ди	0,002521385	ес	0,006764936	жн	0,000782499	зо	0,000862755	ис	0,005373827	йк	0,000842691	ку	0,002711994	ль	0,004169983	мн
7	ар	0,005741668	ба	0,001153684	вы	0,003150059	ги	0,000769123	ду	0,00177567	ем	0,00507621	жу	0,000247457	зе	0,000698898	ик	0,004240207	йв	0,000795875	кн	0,001277413	лс	0,003056427	му
8	аэ	0,0055009	би	0,000842691	вн	0,001585062	гу	0,000578514	дв	0,001257348	ев	0,004537824	жк	0,000197297	зи	0,000698898	им	0,003564717	йи	0,000702242	кн	0,001203844	лу	0,001668662	мс
9	ав	0,005029394	бл	0,000635362	вп	0,001153684	ге	0,000568482	др	0,00125066	ед	0,004036222	жо	8,02563E-05	зу	0,000521666	из	0,003253724	йм	0,000521666	ке	0,00118378	ля	0,001631878	мт
10	ам	0,004611392	ба	0,00042469	вш	0,001140308	гн	0,000545074	дь	0,000705587	ег	0,00402619	жс	5,35042E-05	зл	0,000484882	ич	0,002635081	йд	0,000504946	кт	0,0010333	лн	0,001464677	мф
11	ап	0,003444332	бн	0,000397937	вк	0,001063396	гк	0,000107008	дс	0,000611954	еп	0,003644973	жл	4,68162E-05	зы	0,000481538	ип	0,002534761	йч	0,00046147	кл	0,00101658	лв	0,001140308	мь
12	ад	0,00338414	бь	0,000220705	вр	0,001053364	гп	9,02883E-05	ды	0,000568482	ек	0,002508009	жб	4,68162E-05	зр	0,000394593	ид	0,002444473	йо	0,00045813	кс	0,000628674	лк	0,00110018	ма
13	ая	0,002762154	бк	0,000167201	вт	0,000956387	гс	8,02563E-05	дл	0,000565138	ез	0,002374248	жч	3,67841E-05	зм	0,000337745	ие	0,0021502	йг	0,000384561	кп	0,000444754	лп	0,001080116	мк
14	аи	0,001859271	бщ	0,000157169	вл	0,000919603	гч	6,35362E-05	дя	0,00050829	ей	0,002203704	жв	3,09061E-05	зг	0,000327713	ий	0,001949559	йб	0,000347777	кб	0,000384561	лы	0,000986484	мч
15	ае	0,00164191	бс	0,000123728	ву	0,000906227	гз	4,68162E-05	дк	0,00045813	еб	0,001939527	жм	2,67521E-05	зк	0,000317681	их	0,001842551	йб	0,000331057	кк	0,000334401	лю	0,000812595	мт
16	аг	0,001631878	бм	8,02563E-05	вз	0,000655426	гч	4,68162E-05	дт	0,000337745	еч	0,001768982	жз	2,00641E-05	зп	0,000260833	ио	0,001735542	йш	0,000314337	кч	0,000234081	лт	0,000578514	мл
17	ач	0,001581718	бз	5,35042E-05	вь	0,00062533	гт	4,01281E-05	дп	0,000247457	ее	0,001745574	жв	2,00641E-05	зб	0,000234081	ир	0,001722166	йз	0,000264177	км	0,000190609	лп	0,000565138	мб
18	аж	0,00157503	бя	5,01602E-05	вд	0,000601922	гб	2,67521E-05	дл	0,000247457	еи	0,001461333	жг	1,67201E-05	зя	0,000234081	ии	0,001705446	йу	0,000214017	кж	0,000173889	лч	0,000545074	мт
19	ах	0,001478053	бх	4,34722E-05	вв	0,000585202	гм	1,67201E-05	дм	0,000244113	ео	0,001374389	жп	1,3376E-05	зс	0,000203985	иб	0,001411173	йл	0,000210673	кд	0,000127072	лб	0,00049157	мд
20	аб	0,001414517	бг	2,34081E-05	вм	0,000468162	гг	1,3376E-05	дш	0,000193953	еж	0,001056708	жя	1,3376E-05	зз	0,000143793	ия	0,001407829	йа	0,000183921	кз	0,000120384	лм	0,000411313	мн
21	ао	0,001170404	бж	1,3376E-05	вг	0,000407969	гх	6,68802E-06	дч	0,000123728	еш	0,000953043	жт	1,0032E-05	зь	0,00013376	иц	0,00113362	йж	0,000143793	кэ	0,000103664	лг	0,000381217	мр
22	аш	0,001023268	бп	1,0032E-05	вя	0,000394593	гг	6,68802E-06	дб	0,000107008	еу	0,000749059	жю	1,0032E-05	зж	0,000123728	иг	0,00113362	йц	0,000137104	кг	0,000103664	лд	0,000321025	мэ
23	аю	0,000932979	бч	1,0032E-05	вч	0,000377873	гя	3,34401E-06	дз	0,000103664	ещ	0,000682178	жр	1,0032E-05	зт	7,02242E-05	иш	0,000722307	йя	0,00011704	кх	5,68482E-05	лр	0,000250801	мэ
24	ай	0,000685522	бт	1,0032E-05	вэ	0,000334401			дг	8,69443E-05	ех	0,000648738	жэ	6,68802E-06	зч	5,68482E-05	иу	0,00062533	йе	8,69443E-05	кя	5,35042E-05	лж	0,000244113	мю
25	ау	0,000665458	бб	1,0032E-05	вб	0,000300961			дп	8,02563E-05	ea	0,000434722	жх	6,68802E-06	зэ	3,00961E-05	иа	0,000468162	йэ	8,69443E-05	кш	4,68162E-05	лз	0,000227393	мх
26	аа	0,000444754	бш	6,68802E-06	вц	7,69123E-05			дх	6,68802E-05	ец	0,000428034			зш	2,00641E-05	иж	0,000407969	йх	8,02563E-05	кц	4,01281E-05	лэ	7,02242E-05	мш
27	ащ	0,000434722	бд	6,68802E-06	вх	7,69123E-05			дж	6,35362E-05	ея	0,000407969			зц	1,0032E-05	ию	0,000310993	йф	8,02563E-05	кф	1,0032E-05	лш	7,02242E-05	мы
28	аэ	0,000384561	бю	3,34401E-06	вж	7,35683E-05			дэ	3,00961E-05	еэ	0,000300961			зх	6,68802E-06	ищ	0,000260833	йщ	2,67521E-05	кю	3,34401E-06	лф	6,35362E-05	мэ
29	аф	0,000310993	бц	3,34401E-06	вщ	3,67841E-05			дю	2,00641E-05	еф	0,000234081			зю	3,34401E-06	из	0,000214017	йю	1,0032E-05			лх	4,68162E-05	мь

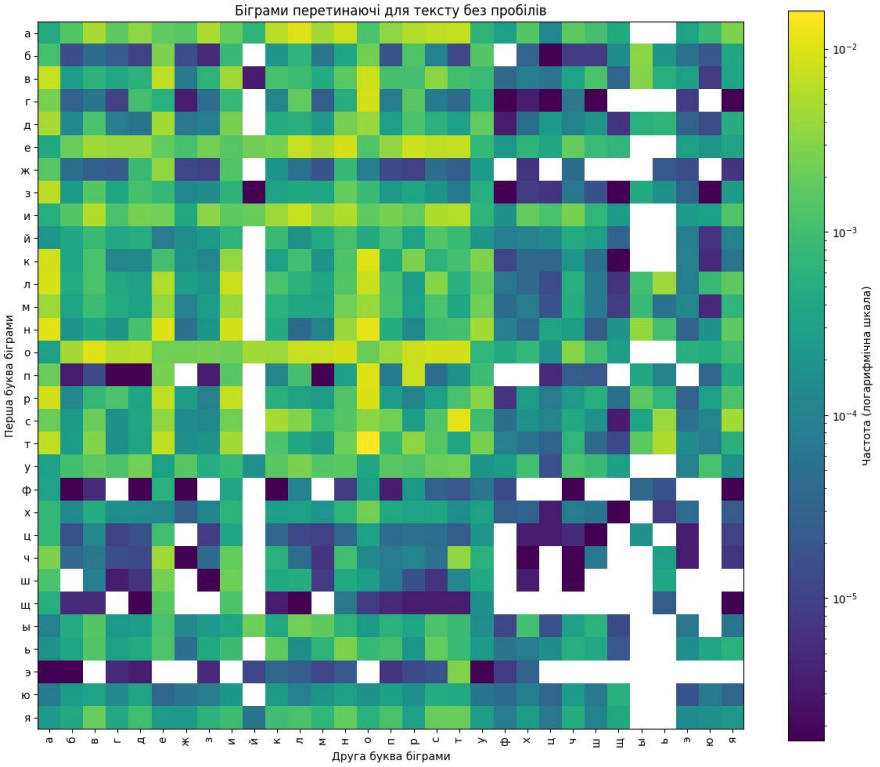
Логування про збереження результатів в excel:

```
[SAVE] Збереження даних [SAVE]

В папку "output\без_пробілів" збережено:
+] Файл частоти_букв_без_пробілів.xlsx
+] Файл частоти_біграм_перетинаючі_без_пробілів.xlsx
+] Файл частоти_біграм_не_перетинаючі_без_пробілів.xlsx
+] Файл сортовані_біграми_перетинаючі_без_пробілів.xlsx
+] Файл сортовані_біграми_не_перетинаючі_без_пробілів.xlsx
```

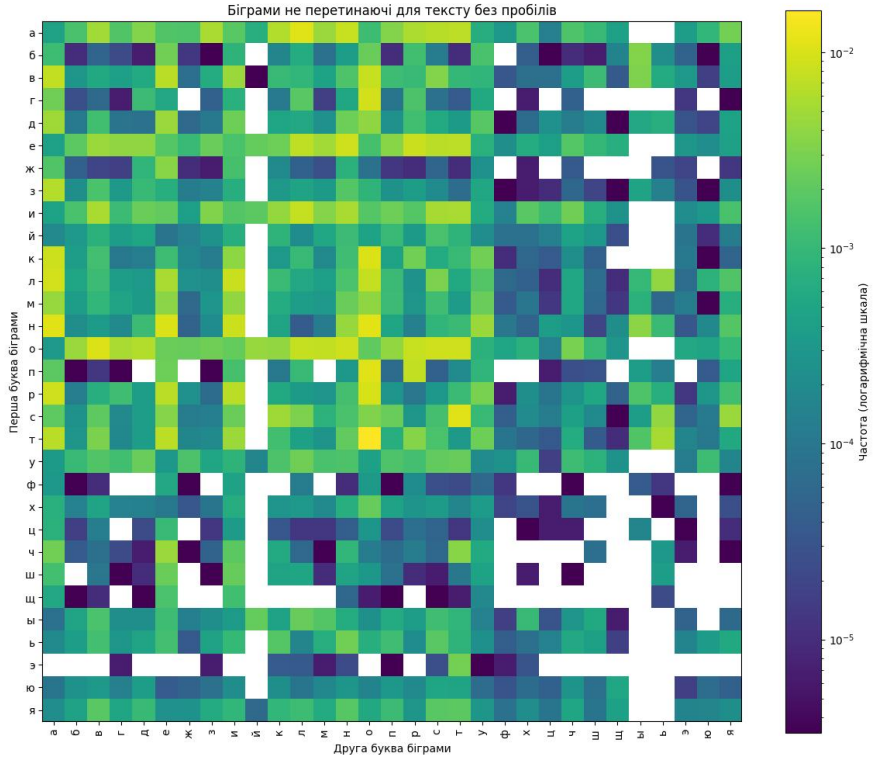
➢ Матриці біграм

```
В папку "output\біграм_матриці" збережено:
+] Файл матриця_біграми_перетинаючі_для_тексту_без_пробілів.png
+] Файл матриця_біграми_не_перетинаючі_для_тексту_без_пробілів.png
```

Матриця біграми перетинаючі для тексту без пробілів



Матриця біграми не перетинаючі для тексту без пробілів



Загальні результати:

```
!=!=!=! РЕЗУЛЬТАТИ АНАЛІЗУ !=!=!=!

[!] Результати аналізу для тексту з пробілами [!]:
Ентропія H_1: 4.37244
Ентропія H_2 (перетинаючі біграми): 3.99650
Ентропія H_2 (не перетинаючі біграми): 3.99642
Надлишковість R_1: 0.12551
Надлишковість R_2 (перетинаючі біграми): 0.20070
Надлишковість R_2 (не перетинаючі біграми): 0.20072

[!] Результати аналізу для тексту без пробілів [!]:
Ентропія H_1: 4.45075
Ентропія H_2 (перетинаючі біграми): 4.14673
Ентропія H_2 (не перетинаючі біграми): 4.14582
Надлишковість R_1: 0.10162
Надлишковість R_2 (перетинаючі біграми): 0.16299
Надлишковість R_2 (не перетинаючі біграми): 0.16317

+] Результати аналізу збережено у файл: d:\cr\crypro-24-25\lab1\redko-shpak_fb-21_seryakov_fb-21_cp1\output\результати_аналізу.xlsx
+] Текст без пробілів збережено у файл: d:\cr\crypro-24-25\lab1\redko-shpak_fb-21_seryakov_fb-21_cp1\output\text_no_spaces.txt
PS D:\cr\crypro-24-25>
```

А також у вигляді таблиці (результати_аналізу.xlsx):

Текст	Параметр	Значення
З пробілами	Ентропія H_1	4,372436188
З пробілами	Ентропія H_2 (перетинаючі біграми)	3,996502021
З пробілами	Ентропія H_2 (не перетинаючі біграми)	3,996419803
З пробілами	Надлишковість R_1	0,125512762
З пробілами	Надлишковість R_2 (перетинаючі біграми)	0,200699596
З пробілами	Надлишковість R_2 (не перетинаючі біграми)	0,200716039
Без пробілів	Ентропія H_1	4,450749542
Без пробілів	Ентропія H_2 (перетинаючі біграми)	4,146728289
Без пробілів	Ентропія H_2 (не перетинаючі біграми)	4,145820901
Без пробілів	Надлишковість R_1	0,101620270
Без пробілів	Надлишковість R_2 (перетинаючі біграми)	0,162986683
Без пробілів	Надлишковість R_2 (не перетинаючі біграми)	0,163169838

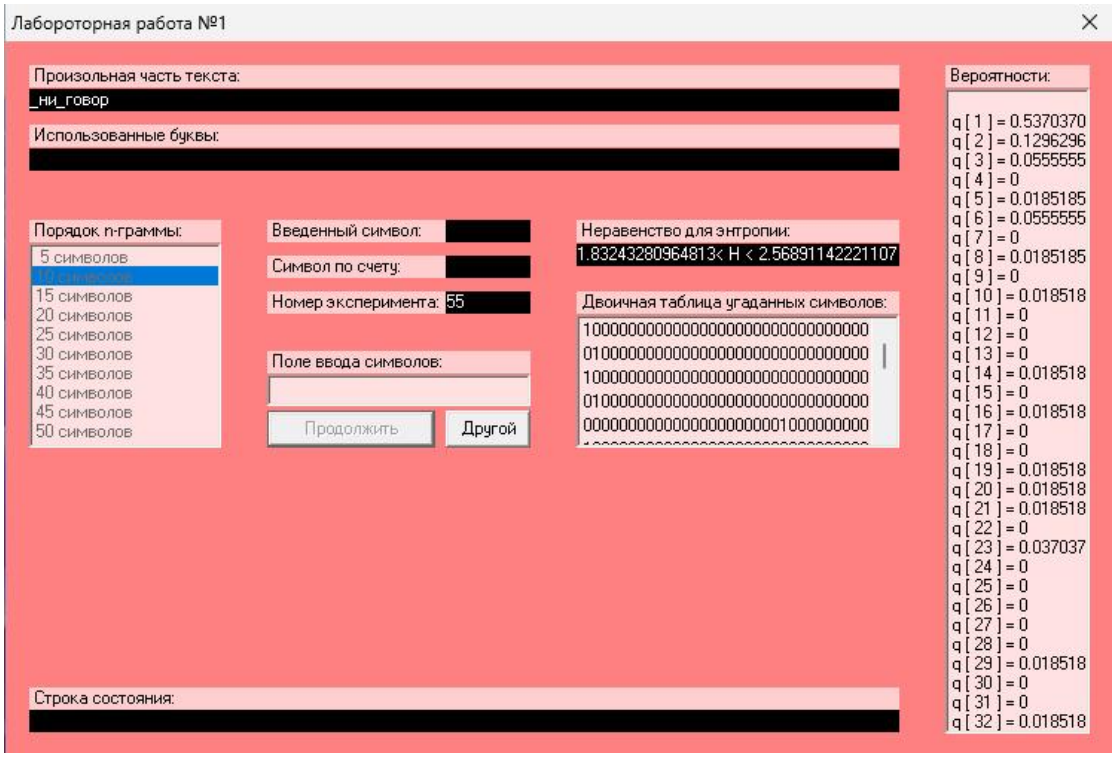
У нас було декілька труднощів:

- Виникли складнощі з коректним обчисленням частот біграм. Довелося розібратися з різницею між перетинаючими та неперетинаючими біграмами.
- Візуалізація матриць біграм спочатку давала некоректні результати. Довелося юзати логарифмічну шкалу для кольорів.

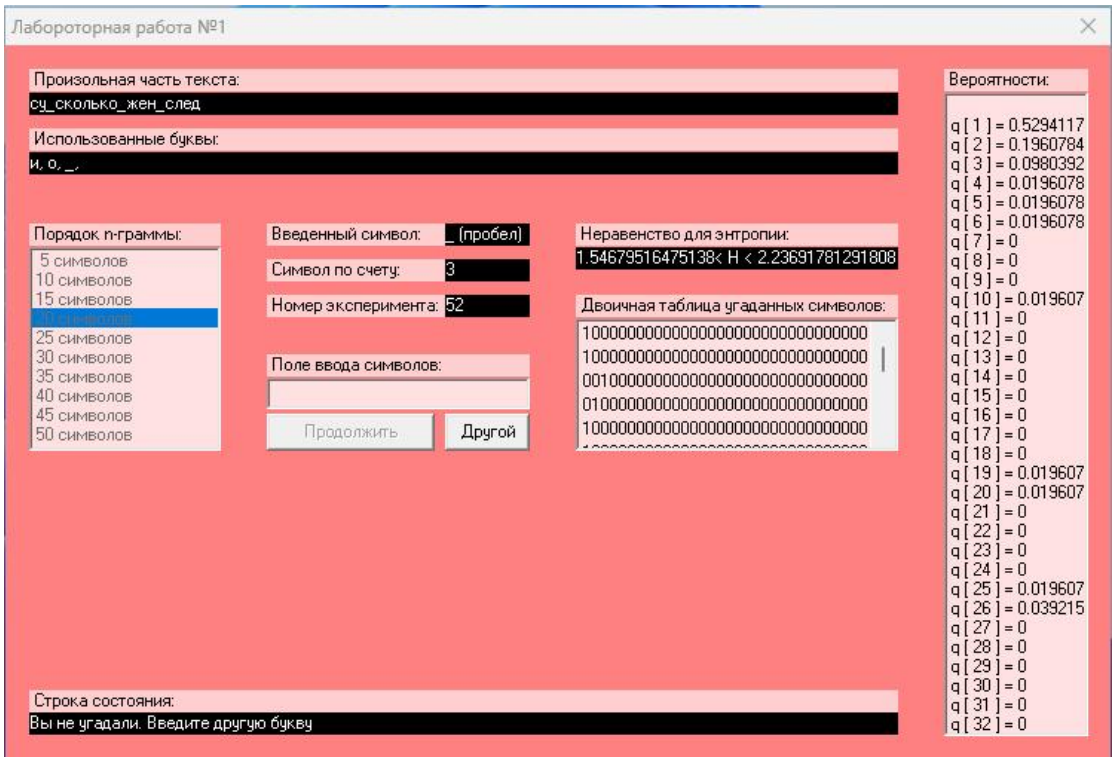
Пункт №2 “За допомогою програми CoolPinkProgram оцінити значення H (10) , H (20) , H (30) .”

Ми ознайомилися з CoolPinkProgram та провели 50+ експериментів для відповідних значень.

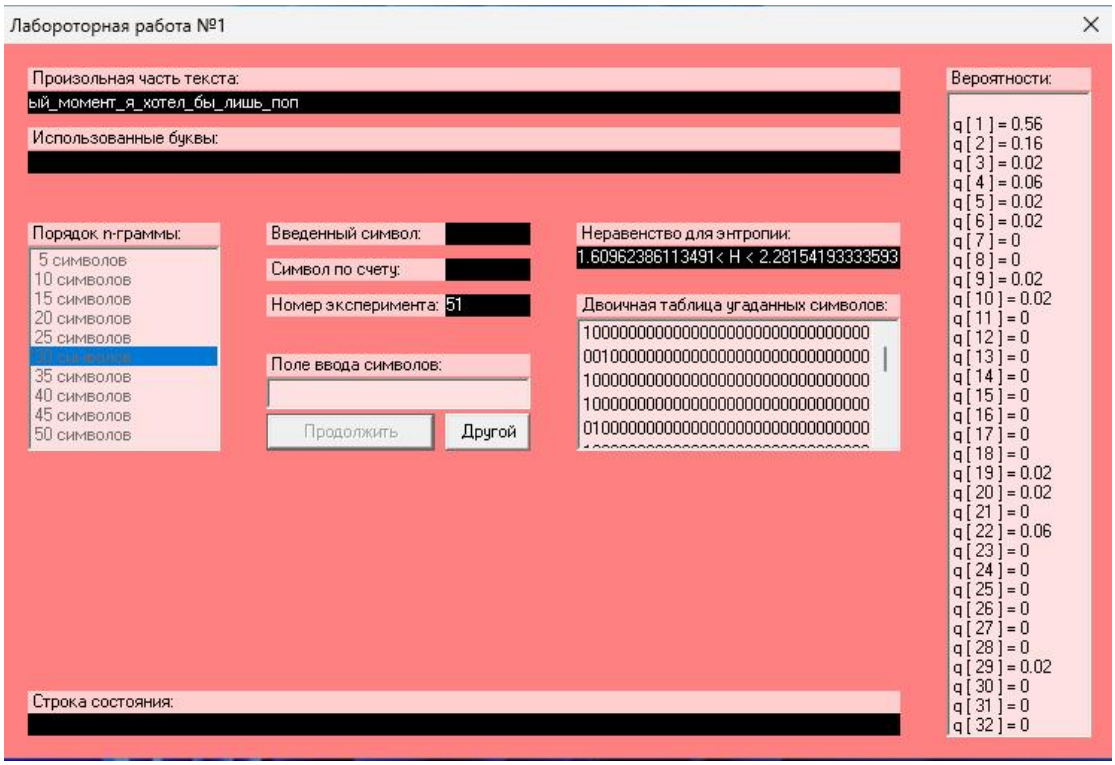
- **H^{10} (55 експериментів):**



- **H^{20} (53 експерименти):**



- **H^{30} (51 експеримент):**



- **Значення ентропії з експериментів:**

H	Мінімальне значення	Максимальне значення
H^{10}	1,8324329095481	2,5889114222110
H^{20}	1,5467951647513	2,2369178129180
H^{30}	1,6095238611349	2,2815419333359

- **Надлишковість для цих експериментів:**

R	Мінімальне значення	Максимальне значення
R для H^{10}	0,63351341809038	0,48221771555780
R для H^{20}	0,69064096704974	0,55261643741640
R для H^{30}	0,67809522777302	0,54369161333282

Пункт №3 Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Аналіз результатів:

Пункт№1 Ентропія H1 для тексту з пробілами (4.37244) виявилася меншою, ніж для тексту без пробілів (4.45075). Це можна пояснити тим, що пробіл є найчастішим символом, що знижує загальну ентропію.

Ентропія H2 для обох варіантів тексту трохи менша за H1, що відповідає теоретичним очікуванням.

Надлишковість тексту з пробілами (20.07%) трохи вища, ніж тексту без пробілів (16.3%). Це вказує на те, що пробіли додають певну надлишковість до тексту.

Матриці біграм демонструють чіткі патерни частот появи пар літер, що відображає структуру мови.

Пункт№2 За допомогою програми CoolPinkProgram ми змогли оцінити ентропію для довших послідовностей символів (H(10), H(20), H(30)). Зі збільшенням довжини n-грами спостерігається зменшення діапазону ентропії, а це вказує на зростання передбачуваності тексту при розгляді довших послідовностей. Значення для довших n-грам є значно нижчими порівняно з H1 та H2.

Висновки

Розроблений пайтон код дозволяє активно аналізувати статистичні характеристики тексту і обчислювати частоти букв та біграм, їх ентропію та надлишковість.

Ми бачимо, що значення ентропії зменшуються при збільшенні порядку (від H1 до H2), що відповідає теоретичним очікуванням. Пробіли у тексті зменшують загальну ентропію, але збільшують надлишковість.

Експериментальні оцінки H^(10), H^(20), H^(30) дають нижчі значення, ніж H1 та H2, що може вказувати на наявність довгострокових залежностей у тексті, які не враховуються при розрахунку H1 та H2.

Я побачив, шо H^(20) і H^(30) дають схожі результати. Можливо, врахування більш ніж 20 попередніх символів не дає істотного покращення в прогнозуванні наступного символу.

За допомогою матриць біграми (хітмап) ми змогли наочно побачити частоти біграм в тексті з пробілами та без.