

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)** H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 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 H_1 та H_2
- ✧ 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

[х] Частоти букв [х]:
: 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

[х/х] Біграми (перетинаючі) [х/х]:
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

[х|х] Біграми (не перетинаючі) [х|х]:
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):

```
[х] Частоти букв [х]:
: 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):

```
[х/х] Біграми (перетинаючі) [х/х]:
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):

```
[х|х] Біграми (не перетинаючі) [х|х]:
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 стовпчику пишеться частота, що відповідає цій біграмі.

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

Сортовані біграми перетинаючі з пробілами

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
	Частота	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	в	х	я	
1	п	0,016614435	а	0,017504619	б	0,002736545	в	0,006776072	г	0,007598861	д	0,004185551	е	0,015756544	ж	0,002955581	з	0,005459049	и	0,015971368	й	0,007779987	ю	0,008401993	я
2	в	0,016002258	ал	0,008633665	б	0,002313918	во	0,006295878	га	0,002134196	де	0,003849977	ер	0,006477004	же	0,00131562	з	0,001558525	ил	0,006130197	ин	0,000405778	ля	0,007070928	ла
3	и	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	0,000491478	йт	0,000391737	к	0,004163086	ли
4	с	0,014571712	ат	0,004917075	бу	0,001213122	ве	0,005610889	гр	0,001277709	ди	0,001982556	кл	0,005919586	жд	0,000716079	ив	0,001092371	ин	0,003260265	ис	0,000376292	лю	0,002883973	ло
5	н	0,010861354	ар	0,004393657	бв	0,001068502	ви	0,003534232	гд	0,000290013	ди	0,001017797	ет	0,005203507	жи	0,00063043	ид	0,000862103	ис	0,002396438	иш	0,000374105	ю	0,002179126	лю
6	о	0,009455005	ак	0,004267956	бо	0,000951964	вы	0,002604562	гн	0,000630343	ду	0,001454623	ж	0,004073225	жу	0,00025835	ю	0,000647279	ив	0,000278071	иш	0,000171297	р	0,002103306	ль
7	к	0,009431181	ас	0,004045144	би	0,000659916	св	0,0022115943	г	0,000494235	д	0,001433562	ем	0,003663235	жу	0,000185338	зе	0,000627622	ик	0,002678978	иш	0,000139004	е	0,000864911	лс
8	т	0,007218356	аз	0,003876655	бл	0,000567247	вш	0,000951964	гу	0,000481598	дв	0,000963196	ет	0,002938732	жи	0,000140408	зи	0,000492831	им	0,002544186	ид	0,000102498	л	0,00080594	лв
9	б	0,006287454	ам	0,003114241	ба	0,000322938	вн	0,00088176	е	0,00047879	дв	0,000918266	ед	0,002813769	им	5,61631E-05	зу	0,000425435	из	0,000218895	иц	8,28405E-05	кт	0,00070625	лу
10	и	0,005749693	ав	0,002497852	бн	0,000315917	ву	0,000650087	н	0,000390333	дь	0,000579884	в	0,002115943	жл	4,07182E-05	зы	0,000381909	ич	0,001157581	им	4,91427E-05	н	0,000689401	лы
11	д	0,005277923	ад	0,002249096	б	0,000190954	вр	0,000614985	к	7,58201E-05	ды	0,00048581	ей	0,001851977	жн	3,51019E-05	эл	0,000343999	и	0,001651194	ио	2,24652E-05	нв	0,000546186	лю
12	ч	0,005288373	ав	0,002117347	бв	0,000178319	ва	0,000586904	н	4,91427E-05	дл	0,000419819	и	0,001452129	ио	2,38693E-05	ит	0,000272391	иш	0,001617496	ио	2,10611E-05	нс	0,000143216	лв
13	н	0,004613795	ак	0,000112361	бв	0,000158661	вн	0,000483002	т	2,10611E-05	дл	0,000411394	и	0,001387227	жн	2,10611E-05	пр	0,00025835	ид	0,001489725	иш	1,8253E-05	нл	4,77386E-05	лп
14	з	0,004552015	ав	0,001106412	бц	0,000129175	св	0,000415607	м	1,12326E-05	дс	0,000359444	е	0,001290346	жл	1,8253E-05	зм	0,000237289	их	0,001437774	иш	1,8253E-05	нц	2,52734E-05	лм
15	г	0,00438493	ак	0,001086755	бс	8,14364E-05	вт	0,000290644	с	8,42446E-06	дл	0,000308176	еп	0,001206102	жт	1,40408E-05	за	0,000199379	ир	0,00110922	иш	1,68489E-05	нк	2,52734E-05	лн
16	у	0,004040932	ак	0,000929499	бм	4,49304E-05	нд	0,000270987	в	7,02038E-06	дц	0,000210611	ч	0,001061482	жм	9,82853E-06	ю	0,000171297	ия	0,000991278	ил	1,54448E-05	кз	1,8253E-05	лт
17	р	0,003864018	ак	0,000914054	бх	2,80815E-05	ва	0,000255542	г	1,40408E-06	дт	0,000203591	еб	0,000869123	жк	9,82853E-06	зь	0,00015164	иц	0,000970217	ип	1,40408E-05	кш	1,12326E-05	лп
18	е	0,002810961	ак	0,000902821	бв	2,10611E-05	вк	0,000233077	п	1,40408E-06	дш	0,000160065	ел	0,000758201	жр	4,21223E-06	жк	0,00011373	ио	0,000571459	иш	1,40408E-05	кп	7,02038E-06	лб
19	а	0,002775859	аш	0,000791899	бт	1,8253E-05	вп	0,000259556	ш	1,40408E-06	дш	0,000160065	ел	0,000758201	жр	4,21223E-06	жк	0,00011373	ио	0,000571459	иш	1,40408E-05	кп	7,02038E-06	лб
20	и	0,002773949	ию	0,000791524	бп	1,68489E-05	вм	0,000212655	ш	0,000139002	иц	0,000160065	ел	0,000758749	ж	8,0527E-05	иш	0,000497043	иш	4,21223E-06	ик	5,61631E-06	кш	4,21223E-06	лп
21	з	0,002367273	ай	0,000575671	бн	1,12326E-05	вн	0,000117942	ш	дб	7,02038E-05	ек	0,000397354	ж	зп	4,35264E-05	нр	0,000494235	иш	4,21223E-06	ик	4,21223E-06	кш	4,21223E-06	лп
22	ж	0,001816875	ай	0,000565843	бб	1,12326E-05	вч	0,000108114	ш	дч	6,73957E-05	ек	0,000329598	ж	зч	1,40408E-05	но	0,000417711	ир	2,80815E-06	ик	4,21223E-06	кш	4,21223E-06	лп
23	н	0,001564141	иц	0,000341191	бш	7,02038E-06	вк	3,93141E-05	ш	дш	5,19508E-05	ек	0,000217632	ж	зс	1,26367E-05	ип	0,000321533	иш	2,80815E-06	ик	4,21223E-06	кш	4,21223E-06	лп
24	х	0,001093775	ай	0,000313109	бд	7,02038E-06	вп	3,79101E-05	ш	дх	3,36978E-05	ио	0,000209207	ж	зт	4,21223E-06	ик	0,000247117	иш	1,40408E-06	ик	4,21223E-06	кш	4,21223E-06	лп
25	ш	0,001030592	аф	0,00018955	бв	5,61631E-06	вг	2,24652E-05	ш	дш	1,40408E-05	ио	0,000178318	ж	зш	2,80815E-06	иш	0,000244309	иш	1,40408E-06	ик	4,21223E-06	кш	4,21223E-06	лп
26	ф	0,000716079	ау	8,14364E-05	бт	4,21223E-06	вц	1,68489E-05	ш	дл	1,12326E-05	иш	0,000174105	ж	зш	2,80815E-06	иш	0,000238693	иш	1,40408E-06	ик	4,21223E-06	кш	4,21223E-06	лп
27	ц	0,000354138	ав	7,10338E-05	бв	1,12326E-05	вп	1,12326E-05	ш	дл	9,82853E-06	иш	0,000162873	ж	зш	2,80815E-06	иш	0,000197975	иш	1,40408E-06	ик	4,21223E-06	кш	4,21223E-06	лп
28	щ	0,000137599	ав	5,4759E-05	бп	1,40408E-06	вп	1,40408E-06	ш	дл	9,82853E-06	иш	0,000134791	ж	зш	2,80815E-06	иш	0,00023559	иш	1,40408E-06	ик	4,21223E-06	кш	4,21223E-06	лп
29	к	0,001375571	аш	5,4759E-05	бп	1,40408E-06	вп	1,40408E-06	ш	дл	9,82853E-06	иш	0,000134791	ж	зш	2,80815E-06	иш	0,00023559	иш	1,40408E-06	ик	4,21223E-06	кш	4,21223E-06	лп

Сортовані біграми не перетинаючі з пробілами

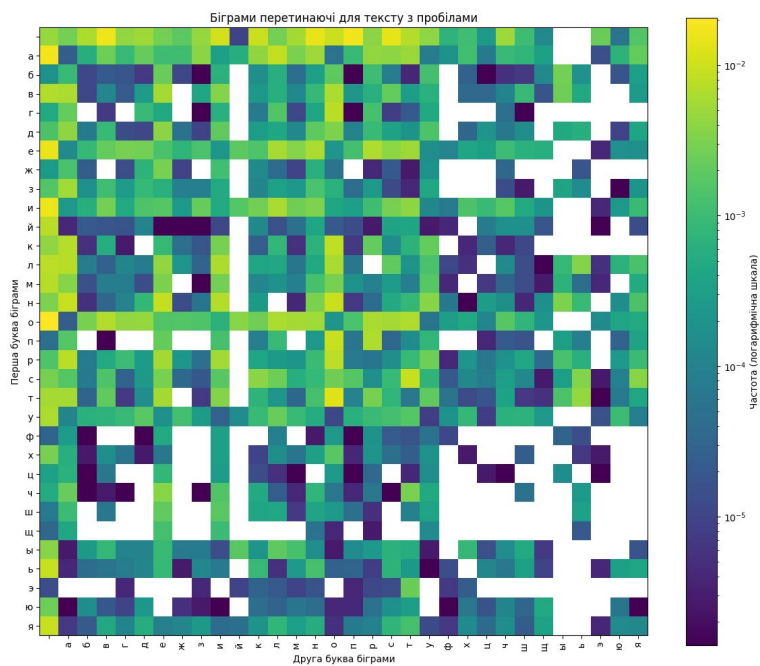
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
1	Частота	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	в	х	я		
2	п	0,016745014	а	0,017390889	б	0,002768838	в	0,006812578	г	0,007438796	д	0,004341404	е	0,016031743	ж	0,002923287	з	0,005459049	и	0,015874487	й	0,007803856	ю	0,008219463	я	
3	в	0,015807901	ал	0,000710899	б	0,002274604	ва	0,006290262	га	0,002252138	де	0,003897716	ер	0,006593542	же	0,00132264	з	0,001555717	ил	0,00622866	йн	0,000387525	к	0,00717483	ла	
4	и	0,014860744	ан	0,001934817	бо	0,001394817	во	0,006256564	гл	0,001536059	до	0,003316428	ен	0,006309919	жи	0,001092371	зн	0,00135353	ит	0,000469013	ис	0,000387525	к	0,004080246	ло	
5	с	0,014661385	ат	0,004917075	бу	0,001221546	ве	0,005512404	гр	0,001235448	ди	0,001917323	ел	0,005922394	жд	0,000693614	ив	0,001165383	ин	0,000373484	иш	0,000296299	л	0,002962999	лю	
6	н	0,010918297	ар	0,00437791	бв	0,001036208	ви	0,003703953	г	0,000848062	ду	0,001861805	ет	0,005242821	жи	0,000634642	ид	0,000859295	ис	0,002920479	иш	0,000232184	л	0,002131388	лю	
7	о	0,009455002	ак	0,004198188	ба	0,000960388	вы	0,002518913	н	0,0006543	д	0,001448634	с	0,003976344	жу	0,000263966	ю	0,000665532	иш	0,000179722	иш	0,000210155	л	0,00210155	лю	
8	к	0,009283753	ас	0,004077438	би	0,000716079	св	0,002137004	г	0,000474578	дв	0,001433939	м	0,003746076	жу	0,000202187	е	0,000617794	ик	0,002693018	иш	0,000162873	л	0,00028405	лс	
9	т	0,007232662	аз	0,003903332	бл	0,000598137	вш	0,000960388	г	0,00044088	дв	0,00092669	ет	0,003049654	ж	0,000143216	и	0,000477386	им	0,002451517	ид	9,82853E-05	н	0,000794797	лш	
10	б	0,00621725	ам	0,003176021	ба	0,000320129	вн	0,000876144	е	0,000432456	дл	0,000915458	ед	0,002900822	жм	6,73957E-05	зу	0,00043264	из	0,000210498	иц	5,89712E-05	кт	0,000687997	лш	
11	и	0,005902737	ав	0,002352567	бн	0,000311705	ву	0,000631834	н	0,000401566	дл	0,000603753	ев	0,00112858	жл	4,77386E-05	зы	0,000426839	иш	0,001631537	им	3,93141E-05	н	0,000682381	лш	
12	д	0,005203507	ад	0,002277859	б	0,000221546	вр	0,000626594	к	6,73957E-05	ды	0,000486961	ей	0,001788369	жн	3,36978E-05	эл	0,001522938	иш	0,001625992	иш	2,52734E-05	нв	0,000682381	лш	
13	ч	0,005192774	ав	0,002052676	бв	0,000179723	ва	0,000556014	н	4,21223E-05	дл	0,000431566	и	0,00152019	ио	2,24652E-05	ит	0,000306989	иш	0,00164688	иш	2,52734E-05	н	0,00071942	лш	
14	ш	0,004580097	ак	0,0014011	к	0,000148832	вб	0,000511084	т	1,96571E-05	дл	0,00039595	н	0,00139846	ж	2,24652E-05	и	0,000280815	иш	0,001437774	ид	2,24652E-05	н	5,33549E-05	лш	
15	щ	0,00436975	к	0,00010975	к	0,000101093	в	0,00012798	т	1,12326E-05	дл	0,00035019	н	1,96571E-05	ж	0,000255482	и	0,000255482	иш	0,001430929	ид	2,24652E-05	н	2,52734E-05	лш	
16	ц	0,004316131	к	0,000118039	к	7,86238E-05	в	0,000030472	т	8,42446E-06	дл	0,000028897	н	0,001173808	ж	1,96571E-05	и	0,000184964	иш	0,001159767	ид	1,68498E-05	н	1,96571E-05	лш	
17	ч	0,004274875	к	0,00102168	к	4,21223E-05	в	0,000238623	т	1,96571E-06	дл	0,000219844	н	0,001058674	ж	1,68498E-05	и	0,000162873	иш	0,001027784	ид	1,68498E-05	н	1,68498E-05	лш	
18	р	0,003808667	к	0,000591964	к	3,65060E-05	вд	0,000261158	т	2,80815E-06	дл	0,000190954	н	0,000904425	жм	1,12326E-05	и	0,00051544	иш	0,000960388	ин	1,12326E-05	н	1,12326E-05	лш	
19	с	0,002805344	к	0,000386609	к	2,80815E-05	в	0,000247117	т	дл	0,000168489	н	0,000749777	ж	2,80815E-06	и	0,000112326	иш	0,0005167	иш	1,40408E-05	н	5,61316E-06	н	1,40408E-05	лш
20	д	0,002729524	к	0,000814364	к	2,24652E-05	в	0,000212844	т	дл	0,000131983	н	0,000665532	ж	з	8,14364E-05	и	0,000474578	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
21	н	0,002664937	к	0,000749777	к	1,68498E-05	вч	0,000143216	т	дл	0,00010671	н	0,000584096	ж	з	7,82015E-05	и	0,000460537	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
22	к	0,002375697	к	0,000603753	к	1,12326E-05	в	0,000137959	т	дл	7,02038E-05	н	0,000418415	ж	з	4,77386E-05	и	0,00044088	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
23	л	0,002356242	к	0,00053266	к	1,00625E-05	в	0,000129175	т	дл	6,73957E-05	н	0,00036746	ж	з	3,149408E-05	и	0,000340182	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
24	м	0,001573053	к	0,0003403	к	5,61316E-06	в	3,93141E-05	т	дл	6,73957E-05	н	0,000216297	ж	з	8,42446E-06	и	0,000304072	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
25	н	0,001120453	к	0,000320129	д	5,61316E-06	в	3,65060E-05	т	дл	2,80815E-05	н	0,000174105	ж	з	5,16313E-06	и	0,000241501	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
26	ш	0,001010935	к	0,000278037	д	5,61316E-06	в	1,68498E-05	т	дл	1,68498E-05	н	0,000171297	ж	з	2,80815E-06	и	0,000224652	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
27	ф	0,000707654	к	9,82853E-05	д	2,80815E-06	в	1,40408E-05	т	дл	1,12326E-05	н	0,000162873	ж	з	2,80815E-06	и	0,000221844	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш	
28	ц	0,00028924	к	7,02038E-05	д	в	5,61316E-06	т	дл	1,12326E-05	н	0,000162873	ж	з	2,80815E-06	и	0,000210611	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш		
29	ч	0,000168489	к	5,33549E-05	д	дл	8,42446E-06	т	дл	1,12326E-05	н	0,000160065	ж	з	2,80815E-06	и	0,000213559	иш	1,40408E-05	н	2,80815E-06	н	2,80815E-06	лш		
30	ш	К	С	Х	Ш	Ч	Щ	Ц	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	
31	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	Ш	Ч	

➤ Матриці біграм

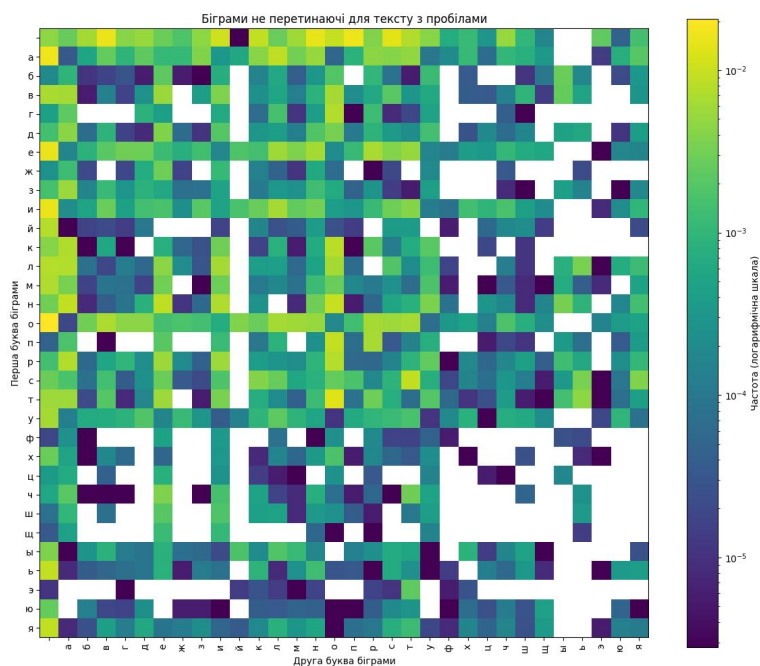
```
В папку "output\біграм_матриці" збережено:  
+) Файл матриця_біграми_перетинаючі_для_тексту_з_пробілами.png  
+) Файл матриця_біграми_не_перетинаючі_для_тексту_з_пробілами.png
```

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 'ь' на 'ъ', 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):

[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

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

[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	еп: 0.0000033440

➤ Таблиці з сортованими біграмами за першими буквами

Сортовані біграми перетинаючі без пробілів (

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
а	Частота а	б	в	г	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
а	0,01068379	б	0,00235874	в	0,00796878	г	0,00908568	д	0,00503738	е	0,00909956	ж	0,00253629	з	0,00642559	и	0,00763605	й	0,00139779	к	0,01025604	л	0,00920940	м
а	0,00817610	б	0,00275881	в	0,00756081	г	0,00255497	д	0,00461975	е	0,00819329	ж	0,00156667	з	0,00184087	и	0,00576507	й	0,00105196	к	0,00805883	л	0,00818469	м
а	0,00687194	б	0,00233746	в	0,00681008	г	0,00181416	д	0,00393423	е	0,00748272	ж	0,00136602	з	0,00147638	и	0,00565472	й	0,00106052	к	0,00382894	л	0,00789019	м
а	0,00677998	б	0,00145631	в	0,00445424	г	0,00154660	д	0,00251804	е	0,00709937	ж	0,00086609	з	0,00111528	и	0,00558156	к	0,00083937	л	0,00271032	м	0,00563632	н
а	0,00637368	б	0,00127409	в	0,00332276	г	0,00109683	д	0,00246953	е	0,00636192	ж	0,00079587	з	0,00087613	и	0,00539389	й	0,00081673	к	0,00261167	л	0,00416329	м
а	0,00567144	б	0,00113694	в	0,00310157	г	0,00081259	д	0,00177064	е	0,00518156	ж	0,00022906	з	0,00076579	и	0,00422850	й	0,00081259	к	0,00127025	л	0,00310829	м
а	0,00521164	б	0,00081259	в	0,00159078	г	0,00059696	д	0,00122558	е	0,00443248	ж	0,00019281	з	0,00065709	и	0,00364330	й	0,00069053	к	0,00117044	л	0,00172551	м
а	0,00569542	б	0,00067549	в	0,00117874	г	0,00058018	д	0,00120718	е	0,00400111	ж	8,3003Е-05	з	0,00053337	и	0,00328047	й	0,00019850	к	0,00163856	л	0,00163856	м
а	0,00458794	б	0,00039615	в	0,00113529	г	0,00059618	д	0,00069538	е	0,00395424	ж	5,1832Е-05	з	0,00045478	и	0,00257651	й	0,00047814	к	0,01050568	л	0,00141781	м
а	0,0034343	б	0,00038289	в	0,00110018	г	0,00011369	д	0,00059021	е	0,00361983	ж	4,6816Е-05	з	0,00043973	и	0,00255984	й	0,00047652	к	0,00103644	л	0,00115034	м
а	0,00339751	б	0,00021234	в	0,00109018	г	7,8584Е-05	д	0,00057854	е	0,00257145	ж	4,5144Е-05	з	0,00038957	и	0,00246119	й	0,00045980	к	0,00065542	л	0,00114868	м
а	0,00272054	б	0,00019727	в	0,00101156	г	7,1896Е-05	д	0,00054507	е	0,00234582	ж	4,3472Е-05	з	0,00038289	и	0,00228396	й	0,00042976	к	0,00043806	л	0,00102824	м
а	0,00184251	б	0,00015382	в	0,00029963	г	6,3536Е-05	д	0,00050327	е	0,00221032	ж	2,6752Е-05	з	0,00036285	и	0,00192651	й	0,00039459	к	0,00037452	л	0,00101992	м
а	0,00174726	б	0,00010536	в	0,00089953	г	6,6848Е-05	д	0,00048482	е	0,00198643	ж	2,5080Е-05	з	0,00033105	и	0,00182081	й	0,00034276	к	0,00030932	л	0,00080759	м
а	0,00165958	б	6,1864Е-05	в	0,00064878	г	4,5144Е-05	д	0,00032697	е	0,00179574	ж	2,1736Е-05	з	0,00026083	и	0,00176294	й	0,00029707	к	0,00021234	л	0,00060526	м
а	0,00137874	б	5,1832Е-05	в	0,00062358	г	4,0128Е-05	д	0,00030961	е	0,00179239	ж	2,1736Е-05	з	0,00025415	и	0,00173218	й	0,00023573	к	0,00019281	л	0,00057182	м
а	0,00149477	б	4,3472Е-05	в	0,00060526	г	3,0961Е-05	д	0,00025080	е	0,00168021	ж	1,8392Е-05	з	0,00024244	и	0,00168203	й	0,00020901	к	0,00018057	л	0,00055106	м
а	0,00145130	б	3,3440Е-05	в	0,00057517	г	2,5080Е-05	д	0,00023408	е	0,00135268	ж	1,5048Е-05	з	0,00019562	и	0,00140485	й	0,00019727	к	0,0001254	л	0,00049157	м
а	0,00142287	б	2,3408Е-05	в	0,00046984	г	1,0032Е-05	д	0,00020231	е	0,00112526	ж	1,3376Е-05	з	0,00018057	и	0,00135432	й	0,00018392	к	0,00011872	л	0,00042301	м
а	0,00119381	б	2,0064Е-05	в	0,00042469	г	8,3600Е-06	д	0,00012707	е	0,00091960	ж	1,1704Е-05	з	0,00015215	и	0,00119046	й	0,00015769	к	0,00011536	л	0,00037452	м
а	0,00110018	б	1,8392Е-05	в	0,00036619	г	3,3440Е-06	д	0,00010536	е	0,00076745	ж	1,0032Е-05	з	0,00011538	и	0,00115702	й	0,00013876	к	9,3632Е-05	л	0,00032105	м
а	0,00088219	б	1,5048Е-05	в	0,00034443	г	3,3440Е-06	д	8,5272Е-05	е	0,00070915	ж	1,0032Е-05	з	7,1896Е-05	и	0,00074069	й	0,00010032	к	6,1864Е-05	л	0,00028927	м
а	0,00080719	б	1,5048Е-05	в	0,00030749	г	1,6720Е-06	д	8,3603Е-06	е	0,00065202	ж	6,6880Е-06	з	6,5208Е-06	и	0,00019032	й	4,8488Е-05	к	4,8488Е-05	л	0,00028897	м
а	0,00065877	б	1,1704Е-05	в	0,00027588	г	1,6720Е-06	д	0,00044636	е	0,00044636	ж	6,6880Е-06	з	3,1788Е-05	и	0,00052501	й	8,6944Е-05	к	4,0128Е-05	л	0,00021903	м
а	0,00047150	б	1,0032Е-05	в	8,0256Е-05	г	1,6720Е-06	д	6,0192Е-05	е	0,00041133	ж	1,6720Е-05	и	0,00044308	й	8,3603Е-05	к	3,5112Е-05	л	8,6944Е-05	м	8,6944Е-05	н
а	0,00043472	б	8,3600Е-06	в	7,1896Е-05	г	1,6720Е-06	д	4,6816Е-05	е	0,00053465	ж	8,6003Е-06	и	0,00029473	й	8,0256Е-05	к	1,1704Е-05	л	7,0224Е-05	м	7,0224Е-05	н
а	0,00034949	б	8,3600Е-06	в	6,1864Е-05	г	1,6720Е-06	д	3,0096Е-05	е	0,00029521	ж	6,6880Е-06	и	0,00025415	й	3,0096Е-05	к	5,0160Е-06	л	6,3536Е-05	м	6,3536Е-05	н
а	0,00030932	б	5,0160Е-06	в	4,0128Е-05	г	1,6720Е-06	д	1,5048Е-05	е	0,00022572	ж	1,6720Е-06	и	0,00024413	й	6,6880Е-06	к	1,6720Е-06	л	4,0128Е-05	м	4,0128Е-05	н

Сортовані біграми не перетинаючі без пробілів

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш
а	0,01082456	б	0,00380796	в	0,00719629	г	0,00912912	д	0,00491235	е	0,00896526	ж	0,00363491	з	0,00643975	и	0,00779154	й	0,00142120	к	0,01055374	л	0,00923905	м	
б	0,00796292	б	0,00265517	в	0,00764106	г	0,00261836	д	0,00460470	е	0,00835686	ж	0,00161850	з	0,00174891	и	0,00557124	й	0,00113027	к	0,00848375	л	0,00824967	м	
в	0,00706253	б	0,00236756	в	0,00683850	г	0,00188602	д	0,00397974	е	0,00734016	ж	0,00138442	з	0,00146133	и	0,00556443	й	0,00106674	к	0,00382220	л	0,00788136	м	
г	0,00665458	б	0,00142789	в	0,00450438	г	0,00153824	д	0,00252807	е	0,00713612	ж	0,00089953	з	0,00101323	и	0,00552094	й	0,00085273	к	0,00272305	л	0,00553434	м	
д	0,00623926	б	0,00124068	в	0,00334068	г	0,00112692	д	0,00252138	е	0,00676493	ж	0,00078249	з	0,00086275	и	0,00537382	й	0,00084261	к	0,00271194	л	0,00416993	м	
е	0,00574168	б	0,00115368	в	0,00150059	г	0,00076912	д	0,00177567	е	0,00507621	ж	0,00024757	з	0,00069898	и	0,00424020	й	0,00079587	к	0,00127743	л	0,00305642	м	
ж	0,00505009	б	0,00084261	в	0,00158502	г	0,00057854	д	0,00125748	е	0,00453782	ж	0,00019529	з	0,00069898	и	0,00356471	й	0,00072242	к	0,00120844	л	0,00166862	м	
з	0,00502934	б	0,00063582	в	0,00115368	г	0,00056842	д	0,00125066	е	0,00403622	ж	8,0256Е-05	з	0,00052166	и	0,00235724	й	0,00052166	к	0,00118378	л	0,00163187	м	
и	0,00461139	б	0,00042469	в	0,00114308	г	0,00054507	д	0,00070587	е	0,00402619	ж	5,3540Е-05	з	0,00044882	и	0,00263581	й	0,00050496	к	0,0010333	л	0,00146467	м	
й	0,00344332	б	0,00037937	в	0,00106396	г	0,00010708	д	0,00061195	е	0,00364493	ж	4,6816Е-05	з	0,00048138	и	0,00253476	й	0,00046147	к	0,00101658	л	0,00114038	м	
к	0,00328414	б	0,00022070	в	0,00105336	г	9,0288Е-05	д	0,00056842	е	0,00250809	ж	4,6816Е-05	з	0,00039459	и	0,00244473	й	0,00045813	к	0,00028674	л	0,00110018	м	
л	0,002762154	б	0,000167201	в	0,000959837	г	8,0256Е-05	д	0,00056138	е	0,00237428	ж	3,67841Е-05	з	0,000337745	и	0,00215052	й	0,000384561	к	0,00044754	л	0,00108016	м	
м	0,00272171	б	0,000157169	в	0,000919603	г	6,3562Е-05	д	0,00050829	е	0,002203704	ж	3,0996Е-05	з	0,000327713	и	0,001995459	й	0,00037777	к	0,00048561	л	0,00096864	м	
н	0,00154915	б	0,00009278	в	0,000558127	г	4,6816Е-05	д	0,000358127	е	0,00199537	ж	2,6751Е-05	з	0,000282551	и	0,00182451	й	0,00034401	к	0,00029591	л	0,00025951	м	
о	0,00163187	б	0,00008538	в	0,00055462	г	4,6816Е-05	д	0,000337745	е	0,00176892	ж	2,0641Е-05	з	0,00026083	и	0,00172552	й	0,000314337	к	0,00023491	л	0,000578514	м	
п	0,001581718	б	5,35042Е-05	в	0,00052633	г	4,01281Е-05	д	0,000247457	е	0,001745574	ж	2,0641Е-05	з	0,00023491	и	0,00172564	й	0,000264717	к	0,000190609	л	0,000555138	м	
р	0,00157503	б	5,01602Е-05	в	0,000601922	г	6,67521Е-05	д	0,000247457	е	0,00146133	ж	1,3776Е-05	з	0,00023491	и	0,00175443	й	0,000214017	к	0,00017389	л	0,000540574	м	
с	0,001478053	б	4,34722Е-05	в	0,000585202	г	1,67201Е-05	д	0,000244113	е	0,001374389	ж	1,3776Е-05	з	0,00023491	и	0,00170516	й	0,000210673	к	0,000127072	л	0,00049157	м	
т	0,00144517	б	2,34081Е-05	в	0,000468162	г	1,3776Е-05	д	0,000193953	е	0,00105678	ж	1,1337Е-05	з	0,000143793	и	0,00147829	й	0,000183921	к	0,000120384	л	0,000411313	м	
у	0,001170404	б	1,3776Е-05	в	0,000407969	г	6,68802Е-06	д	0,000123728	е	0,000590340	ж	1,0037Е-05	з	0,000143793	и	0,00113826	й	0,000143793	к	0,000103664	л	0,000381217	м	
ф	0,000123268	б	1,0037Е-05	в	0,000394593	г	6,68802Е-06	д	0,00010708	е	0,000749509	ж	1,0037Е-05	з	0,000123728	и	0,00113362	й	0,000137014	к	0,000103664	л	0,000321025	м	
х	0,0000937873	б	1,0037Е-05	в	0,000394593	г	6,68802Е-06	д	0,000081738	е	0,000681738	ж	1,0037Е-05	з	0,000123728	и	0,00113362	й	0,000137014	к	0,000103664	л	0,000321025	м	
ц	0,000085527	б	1,0037Е-05	в	0,000334401	г	6,68802Е-06	д	8,6944Е-05	е	0,000648738	ж	6,68802Е-06	з	0,000123728	и	0,000682105	й	0,000062533	к	8,6944Е-05	л	0,000244113	м	
ч	0,000065458	б	1,0037Е-05	в	0,00030961	г	8,0256Е-05	д	0,00044722	е	0,00044722	ж	6,68802Е-06	з	0,0006105	и	0,000468162	й	8,6944Е-05	к	4,6816Е-05	л	0,00027393	м	
ш	0,000444754	б	6,68802Е-06	в	7,69123Е-05	г	6,68802Е-05	д	0,000428034	е	3,00611Е-05	ж	0,000417969	з	0,0006105	и	0,000407969	й	8,0256Е-05	к	4,01281Е-05	л	7,02242Е-05	м	
щ	0,000344722	б	6,68802Е-06	в	7,69123Е-05	г	6,68802Е-05	д	0,000407969	е	3,63526Е-05	ж	0,000407969	з	0,0006105	и	0,00031993	й	8,0256Е-05	к	1,0037Е-05	л	7,02242Е-05	м	
ъ	0,000384561	б	3,34401Е-06	в	3,34401Е-06	г	3,34401Е-06	д	3,00611Е-05	е	0,00030961	ж	6,68802Е-06	з	0,0006105	и	0,000260833	й	2,67521Е-05	к	3,34401Е-06	л	6,3562Е-05	м	
ы	0,000310093	б	3,34401Е-06	в	3,67841Е-05	г	3,67841Е-05	д	2,00461Е-05	е	0,000234081	ж	3,34401Е-06	з	0,000214017	и	0,000214017	й	1,0037Е-05	к	0,00017901	л	4,6816Е-05	м	
я																									
К	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш
Sheet1																									

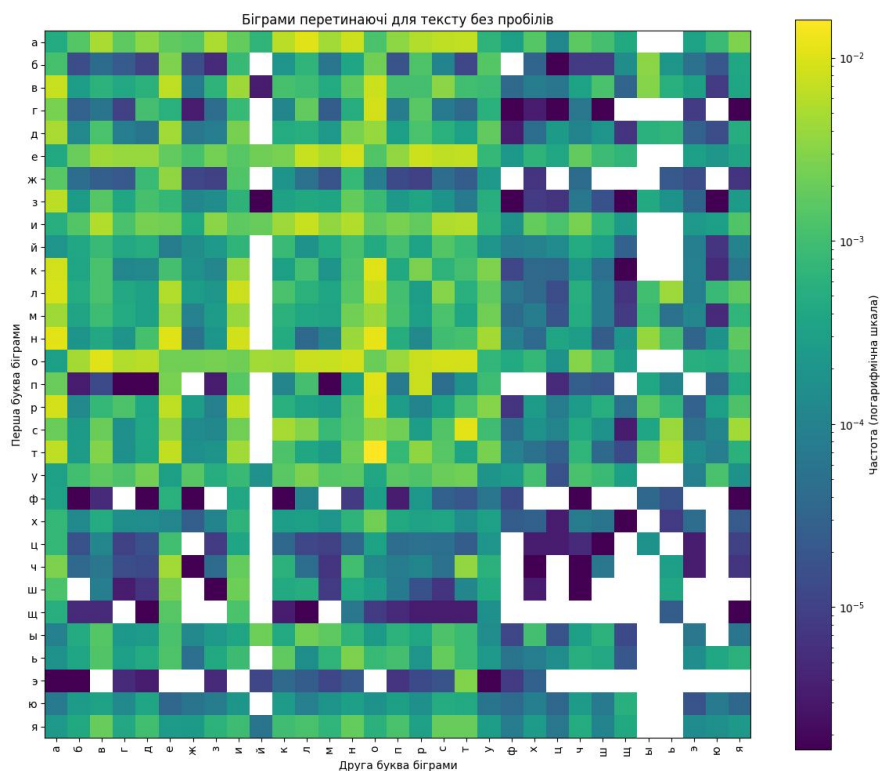
➤ Матриці біграм

В папку "output\біграм_матриці" збережено:

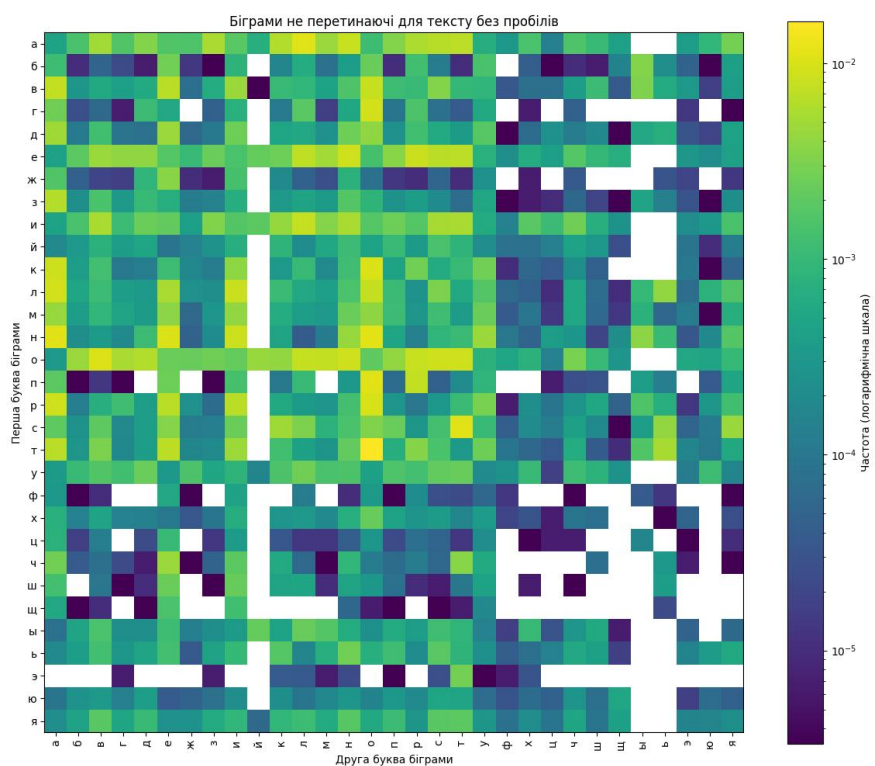
+] Файл матриця_біграми_перетинаючі_для_тексту_без_пробілів.png

+] 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 Н_1: 4.37244  
Ентропія Н_2 (перетинаючі біграми): 7.99300  
Ентропія Н_2 (не перетинаючі біграми): 7.99284  
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 Н_1: 4.45075  
Ентропія Н_2 (перетинаючі біграми): 8.29346  
Ентропія Н_2 (не перетинаючі біграми): 8.29164  
Надлишковість R_1: 0.10985  
Надлишковість R_2 (перетинаючі біграми): 0.17065  
Надлишковість R_2 (не перетинаючі біграми): 0.17084  
  
+] 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 Н_1	4,372436188
З пробілами	Ентропія Н_2 (перетинаючі біграми)	7,993004042
З пробілами	Ентропія Н_2 (не перетинаючі біграми)	7,992839605
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 Н_1	4,450749542
Без пробілів	Ентропія Н_2 (перетинаючі біграми)	8,293456578
Без пробілів	Ентропія Н_2 (не перетинаючі біграми)	8,291641801
Без пробілів	Надлишковість R_1	0,109850092
Без пробілів	Надлишковість R_2 (перетинаючі біграми)	0,170654342
Без пробілів	Надлишковість R_2 (не перетинаючі біграми)	0,17083582

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.

➤ **H¹⁰ (55 експериментів):**

[illegible]

➤ **H²⁰ (53 эксперименти):**

[illegible]

➤ **H³⁰ (51 эксперимент):**

[illegible]

➤ **Значення ентропії з експериментів:**

Н	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:**

Р	Мінімальне значення	Максимальне значення
Р для Н ¹⁰	0,63351341809038	0,48221771555780
Р для Н ²⁰	0,69064096704974	0,55261643741640
Р для Н ³⁰	0,67809522777302	0,54369161333282

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

Аналіз результатів:

Пункт №1 Ентропія H_1 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 H_2 для обох варіантів тексту приблизно вдвічі більша за H_1 , що відповідає теоретичним очікуванням.

Надлишковість тексту з пробілами (20.07%) дещо вища, ніж тексту без пробілів (17.07%). 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 довгих послідовностей символів ($H(10)$, $H(20)$, $H(30)$). Зі збільшенням довжини n-грам спостерігається зменшення діапазону ентропії, а це вказує на зростання передбачуваності тексту при розгляді довгих послідовностей. Значення для довгих n-грам є значно нижчими порівняно з H_1 та H_2 .

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 H_1 до H_2), 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 H_1 та H_2 , що може вказувати на наявність довгострокових залежностей у тексті, які не враховуються при розрахунку H_1 та H_2 .

Цікаво те, що $H^{(20)}$ і $H^{(30)}$ дають схожі результати. Можливо, врахування більш ніж 20 попередніх символів не дає істотного покращення в прогнозуванні наступного символу.

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