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

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1 3 дисципліни «Криптографія»

Виконав:

студент 3 курсу

НН ФТІ групи ФБ-25

Черняк Денис

Тема: «Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Мета: 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.

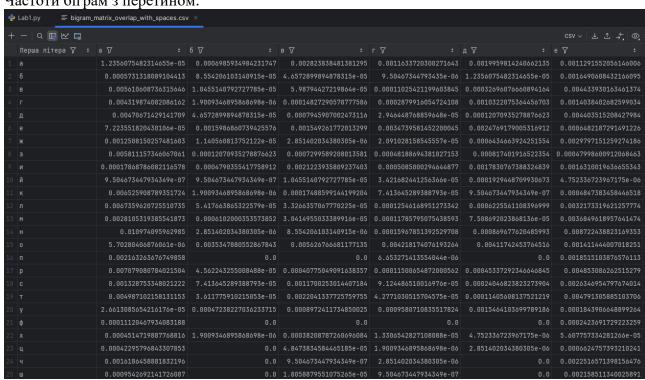
Хід роботи

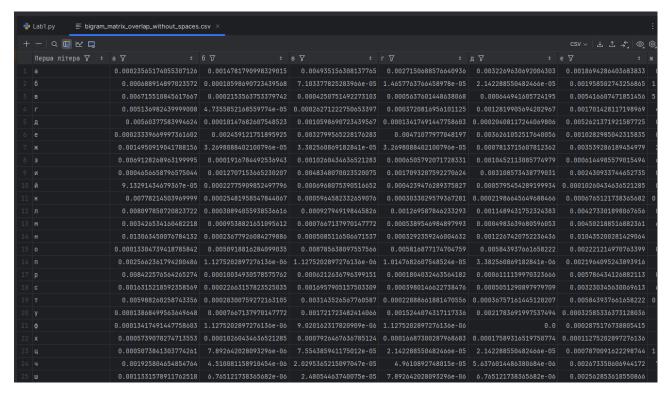
Завдання 1

Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 на тому ж тексті, в якому вилучено всі пробіли.

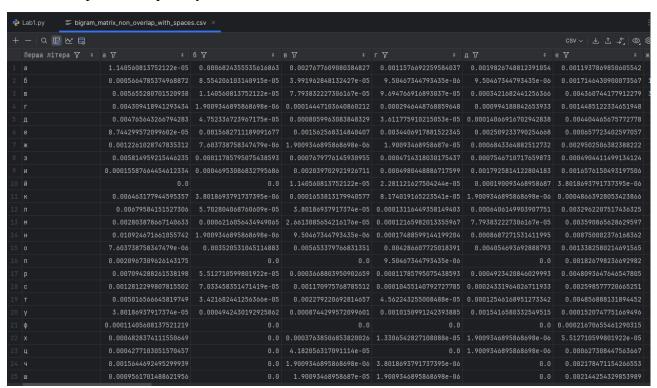
Виконання

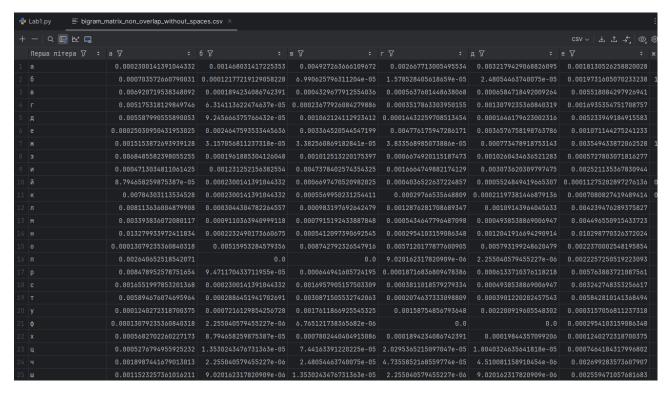
Частоти біграм з перетином:





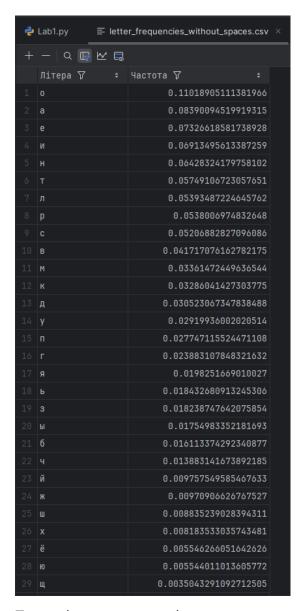
Частоти біграм без перетину:





Частоти букв:

ے.	1 - 1-1			_		u for more describes and the conservation of
~	Lab1	.ру		=	lette	r_frequencies_with_spaces.csv ×
+		Q		₩		
	Літ	ера	V			Частота ▽ ÷
1						0.1570284617175879
2						0.09288623391929589
3	а					0.07072610883791221
4	е					0.06176130936256968
5	И					0.05827880032125766
6	н					0.05418894322388712
7	Т					0.048463333380856656
8	л					0.045465562224661756
9	р					0.045352456718134426
10						0.043892540264134625
11	В					0.035166307865585035
12	м					0.028336256017640657
13	к					0.027700393968339964
14	д					0.025730077035305075
15	у					0.02461422943309429
16	п					0.023390028656563207
17						0.020132780161864434
18	я					0.01671205143924381
19	Ь					0.01553822538410725
20	3					0.015374745156185398
21	Ы					0.01479401016048626
22	б					0.013583115914134861
23						0.011703093293033557
24	Й					0.008225336583928563
25	ж					0.0081844665269481
26	ш					0.007447855034858356
27	х					0.006898485431725619



Ентропія та надлишковість:

3 пробілами:

Тексти	Н	R
Текст з пробілами монограми	4.406696524751244	0.12641708389912865
Текст з пробілами біграми(з перетином)	4.0188120068200375	0.210055752302588
Текст з пробілами біграми(без перетину)	4.018162669281308	0.2101833871490707

Без пробілів:

Тексти	Н	R
Текст без пробілів монограми	4.48359558232663	0.1111726252474392
Текст без пробілів біграми(з перетином)	4.178746528108722	0.17160586004327194
Текст без пробілів біграми(без перетину)	4.178337155941605	0.1716870139256671

Результат у програмі:

```
Ентропія та надлишковість для тексту з пробілами:
Н1 (ентропія букв): 4.406696524751244, R: 0.13381253834018836
Н2 (ентропія біграм з перетином): 4.0188120068200375, R: 0.210055752302588
Н2 (ентропія біграм без перетину): 4.018162669281308, R: 0.2101833871490707

Ентропія та надлишковість для тексту без пробілів:
Н1 (ентропія букв): 4.48359558232663, R: 0.1111726252474392
Н2 (ентропія біграм з перетином): 4.178746528108722, R: 0.17160586004327194
Н2 (ентропія біграм без перетину): 4.178337155941605, R: 0.1716870139256671
```

Завдання 2

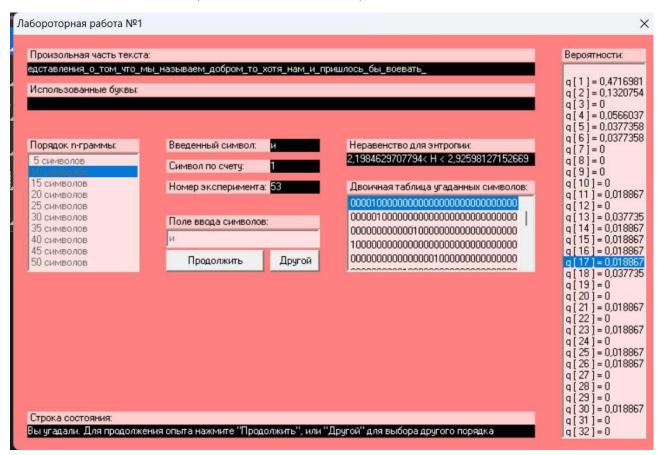
За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}, H^{(20)}, H^{(30)}$.

Під час виконання були проблеми з ієроглєфами, і на жаль, повністю прибрати її не вдалось, але залишок символів не заважав виконання роботи, оскільки одна кнопка відповідала за продовження дії, а інша до перехід до наступного значення

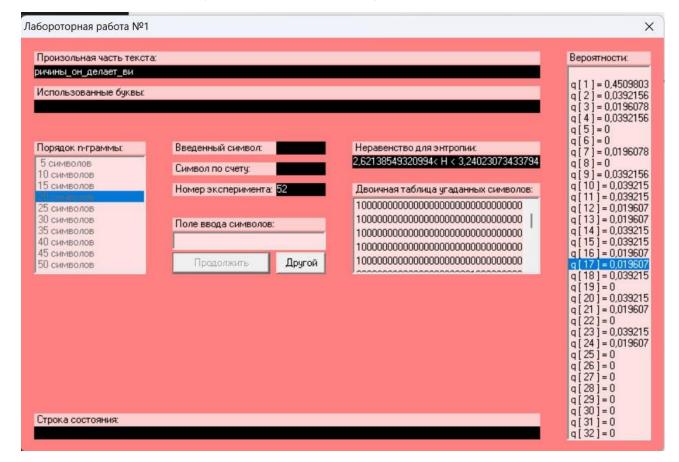
Виконання

 $H^{(10)}$

2,1984629707794 < H < 2,92598127152669

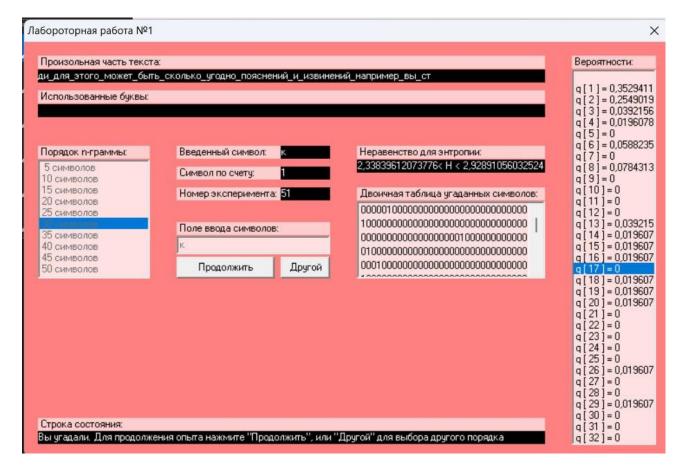


 $H^{(20)}$ 2,62138549320994 < H < 3,24023073433794



 $H^{(30)}$

2,33839612073776 < H < 2,92891056032524



Завдання 3

Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Виконання

Для обрахунків використаємо цю формулу

$$R = 1 - \frac{H_{\infty}}{H_0}$$

 $H_{(0)}$ можна знайти за наступною формулою, 32 це кількість літер у нашому алфавіті, це вказано у методичці, що у наданому тексті використовується лише 32 символи

$$H_{(0)} = \log_2 32 = 5$$

$$H^{(10)}$$
 $R = 1 - \frac{2,1984629707794}{5} \approx 0,56030740584412$
 $R = 1 - \frac{2,92598127152669}{5} \approx 0,414803745694662$
 $H^{(20)}$
 $R = 1 - \frac{2,62138549320994}{5} \approx 0,475722901358012$
 $R = 1 - \frac{3,24023073433794}{5} \approx 0,351953853132412$

$$H^{(30)}$$

$$R = 1 - \frac{2,33839612073776}{5} \approx 0,532320775852448$$

$$R = 1 - \frac{2,92891056032524}{5} \approx 0,414217887934952$$

Висновки:

У ході виконання роботи я дослідив та оцінив ентропію для символів джерела відкритого тексту російською мовою. Було розроблено програму для підрахунку частот символів та біграм у тексті, а також обчислено ентропії Н1 та Н2 (з перетином та без) як з урахуванням пробілів, так і без них. Це дало можливість більш детально вивчити вплив пробілів на загальну ентропію тексту. Зокрема, результати показали, що включення пробілів впливає на розподіл частот окремих символів та біграм, проте не призводить до суттєвих змін значень Н1 та Н2, що свідчить про низький внесок пробілів у загальну ентропію при обробці тексту значного обсягу.

Додатково, за допомогою програми CoolPinkProgram було обчислено ентропії H10, H20, H30 для тексту, що включає лише літери та пробіли, без урахування розділових знаків. Це дозволило оцінити надлишковість російської мови, яка, за нашими підрахунками, варіюється від 47% до 65% залежно від значення ентропії ННН. Високий рівень надлишковості свідчить про значну передбачуваність структури російської мови, що можна враховувати при проєктуванні систем стиснення тексту або криптографічного аналізу.

Результати також вказали на помітну різницю у значеннях Н між текстом із пробілами та без них, а також між різними завданнями, що підтверджує залежність ентропії від обраного тексту та характеру символів, які він містить.