МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»
Фізико-технічний інститут

Криптографія
Комп’ютерний практикум №1

Виконали:

студенти групи ФБ-01

Чуйко О. М.

Ченський К. Ю.

Київ - 2022

ФБ-01 Чуйко Олександр, ФБ-01 Ченський Костянтин

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.

Постановка задачі:

Написати програми для підрахунку частот букв і частот біграм в тексті, також для 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, де імовірності замінити відповідними частотами. Одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

Використавши CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.

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

Хід роботи:

Спершу для знайденого тексту російською мовою провели обробку(перевели усі символи в нижній регістр, видалили повторні пробіли та усі символи крім зазначених), отримали два очищених файли(один з пробілами інший без) та дві відповідні змінні.

Далі написали функцію, яка б рахувала кількість елементів у тексті та сортувала отриманий масив. Після неї написали функцію, яка б виконувала відповідні дії для біграм, при тому з можливістю вказати крок, який для нашої роботи міг бути 1 або 2.

Після цього реалізували функцію, яка обраховує частоту та заносить дані у .xlsx табличку. На цьому етапі виникла проблема: використаний модуль для запису у .xlsx файл міг виконувати дії лише після створення, без можливості подальшого редагування, тож для кожного випадку було необхідно створити окремий файл, але для зручності перевірки та написання протоколу усі отримані файли були об’єднані у один.

Далі написали функції для підрахунку ентропії та надлишковості, використовуючи функції наведені у методичці до комп’ютерного практикуму.

Пілся цього у функції main() зробили 6 викликів для знаходження значення ентропії для усіх випадків, потім 6 викликів для підрахунку надлишковості, після чого занесли відповідні дані у .xlsx табличку.

Таблиці частот букв:

ФБ-01 Чуйко Олександр, ФБ-01 Ченський Костянтин

	A	B
1	Symbol	Frequency
2	' '	0.15709799806323
3	'о'	0.0885243843969633
4	'е'	0.0707176337257712
5	'а'	0.0657611380674483
6	'и'	0.0564049497819127
7	'т'	0.0546900256362916
8	'н'	0.0527758783877962
9	'с'	0.0450703958117904
10	'р'	0.0401598747156692
11	'л'	0.0389455624695783
12	'в'	0.0382989995789314
13	'м'	0.0291704948395378
14	'к'	0.0290982490989351
15	'д'	0.026894389133073
16	'у'	0.0226048895041534
17	'п'	0.022277959284052
18	'я'	0.0180096225488443
19	'ь'	0.0167150372576393
20	'ы'	0.0164202162555835
21	'г'	0.0152372834725827
22	'з'	0.0151103061103112
23	'б'	0.0147191574541185
24	'й'	0.0134209233881355
25	'ч'	0.0125817051892145
26	'ж'	0.00827323193144974
27	'х'	0.00729681980087906
28	'ш'	0.00671228608145671
29	'ю'	0.00533888725504863
30	'ц'	0.00383851106636902
31	'щ'	0.00296280511966887
32	'э'	0.00289639741871077
33	'ф'	0.00173097875464398
34	'ъ'	0.00024300840020929
35	'ё'	0

	A	B
1	Symbol	Frequency
2	'о'	0.105023341021325
3	'е'	0.0838978120389802
4	'а'	0.07801753693353
5	'и'	0.0669175653306183
6	'т'	0.0648830178486089
7	'н'	0.062612116552732
8	'с'	0.0534705051218651
9	'р'	0.0476447732042305
10	'л'	0.0462041404339906
11	'в'	0.0454370727450288
12	'м'	0.034607219786537
13	'к'	0.0345215090628268
14	'д'	0.0319068991072232
15	'у'	0.0268179331075418
16	'п'	0.0264300704386311
17	'я'	0.0213662116206426
18	'ь'	0.019830344712947
19	'ы'	0.0194805756990186
20	'г'	0.0180771708188751
21	'з'	0.0179265277287178
22	'б'	0.0174624777498424
23	'й'	0.015922282017717
24	'ч'	0.0149266523988614
25	'ж'	0.00981517651214495
26	'х'	0.00865678309472853
27	'ш'	0.00796330542107341
28	'ю'	0.00633393590569396
29	'ц'	0.00455392330015722
30	'щ'	0.00351500543700348
31	'э'	0.00343622083238099
32	'ф'	0.00205359430950056
33	'ъ'	0.00028829970702516
34	'ё'	0

Перша табличка демонструє частоти для букв у тексті з пробілами, друга частоти для букв у тексті без пробілів.

Топ-10 по кількості для кожного з 6-ти випадків:

```
[+]Найчастіші 10 елементів:  [(' ', 215275), ('о', 121307), ('е', 96906), ('а', 90114), ('и', 77293), ('т', 74943), ('н', 72320), ('с', 61761), ('р', 55032), ('л', 53368)]
[+]Найчастіші 10 елементів:  [('о', 121307), ('е', 96906), ('а', 90114), ('и', 77293), ('т', 74943), ('н', 72320), ('с', 61761), ('р', 55032), ('л', 53368), ('в', 52482)]
[+]Найчастіші 10 елементів:  [('о ', 23038), (' с', 21736), ('и ', 21663), (' в', 21334), ('е ', 21331), (' п', 20444), ('а ', 19822), (' н', 17788), ('то', 17313), ('ст', 14783)]
[+]Найчастіші 10 елементів:  [('то', 17839), ('ст', 15131), ('на', 12676), ('но', 12066), ('не', 12009), ('ен', 11828), ('по', 10968), ('ко', 10965), ('ов', 10882), ('ра', 10615)]
[+]Найчастіші 10 елементів:  [('о ', 23038), (' с', 21736), ('и ', 21663), (' в', 21334), ('е ', 21331), (' п', 20444), ('а ', 19822), (' н', 17788), ('то', 17313), ('ст', 14783)]
[+]Найчастіші 10 елементів:  [('то', 17839), ('ст', 15131), ('на', 12676), ('но', 12066), ('не', 12009), ('ен', 11828), ('по', 10968), ('ко', 10965), ('ов', 10882), ('ра', 10615)]
```

Таблиці частот біграм:

ФБ-01 Чуйко Олександр, ФБ-01 Ченський Костянтин

	A	B
1	Bigram	Frequency
2	'о '	0.01681
3	' с'	0.01586
4	'и '	0.01581
5	' в'	0.01557
6	'е '	0.01557
7	' п'	0.01492
8	'а '	0.01447
9	' н'	0.01298
10	'то'	0.01263
11	'ст'	0.01079
12	'я '	0.01078
13	'ь '	0.01001
14	'й '	0.00976
15	' и'	0.0093
16	'на'	0.00921
17	' т'	0.00918
18	' к'	0.00904
19	' о'	0.00884
20	'не'	0.00873
21	'но'	0.00862
22	'м '	0.00832
23	'по'	0.008
24	'ко'	0.00773
25	'ра'	0.0077
26	'ен'	0.00742
27	'т '	0.00741
28	' м'	0.00736
29	'ро'	0.00702
30	'во'	0.00689
31	'ни'	0.00675
32	' д'	0.00672
33	'ер'	0.00663
34	'от'	0.00653
35	'го'	0.00651
36	'в '	0.00651
37	'та'	0.00649
38	'ка'	0.00639
39	'ал'	0.00624
40	'ов'	0.0062

	A	B
1	Bigram	Frequency
2	'то'	0.01544
3	'ст'	0.0131
4	'на'	0.01097
5	'но'	0.01045
6	'не'	0.0104
7	'ен'	0.01024
8	'по'	0.0095
9	'ко'	0.00949
10	'ов'	0.00942
11	'ра'	0.00919
12	'от'	0.00893
13	'во'	0.00861
14	'ро'	0.00845
15	'ес'	0.00844
16	'ер'	0.00842
17	'ос'	0.0084
18	'ни'	0.00839
19	'ет'	0.00795
20	'та'	0.00785
21	'го'	0.0078
22	'ал'	0.00774
23	'он'	0.0077
24	'ка'	0.00768
25	'ли'	0.00763
26	'ом'	0.00743
27	'ре'	0.00737
28	'пр'	0.00724
29	'ор'	0.00711
30	'ол'	0.00698
31	'ла'	0.00695
32	'ан'	0.00675
33	'ри'	0.00637
34	'ть'	0.00633
35	'ак'	0.0063
36	'те'	0.00625
37	'ве'	0.00622
38	'ло'	0.00615
39	'ел'	0.00611
40	'од'	0.00611

	A	B
1	Bigram	Frequency
2	'о '	0.01681
3	' с'	0.01586
4	'и '	0.01581
5	' в'	0.01557
6	'е '	0.01557
7	' п'	0.01492
8	'а '	0.01447
9	' н'	0.01298
10	'то'	0.01263
11	'ст'	0.01079
12	'я '	0.01078
13	'ь '	0.01001
14	'й '	0.00976
15	' и'	0.0093
16	'на'	0.00921
17	' т'	0.00918
18	' к'	0.00904
19	' о'	0.00884
20	'не'	0.00873
21	'но'	0.00862
22	'м '	0.00832
23	'по'	0.008
24	'ко'	0.00773
25	'ра'	0.0077
26	'ен'	0.00742
27	'т '	0.00741
28	' м'	0.00736
29	'ро'	0.00702
30	'во'	0.00689
31	'ни'	0.00675
32	' д'	0.00672
33	'ер'	0.00663
34	'от'	0.00653
35	'го'	0.00651
36	'в '	0.00651
37	'та'	0.00649
38	'ка'	0.00639
39	'ал'	0.00624
40	'ов'	0.0062

	A	B
1	Bigram	Frequency
2	'то'	0.01544
3	'ст'	0.0131
4	'на'	0.01097
5	'но'	0.01045
6	'не'	0.0104
7	'ен'	0.01024
8	'по'	0.0095
9	'ко'	0.00949
10	'ов'	0.00942
11	'ра'	0.00919
12	'от'	0.00893
13	'во'	0.00861
14	'ро'	0.00845
15	'ес'	0.00844
16	'ер'	0.00842
17	'ос'	0.0084
18	'ни'	0.00839
19	'ет'	0.00795
20	'та'	0.00785
21	'го'	0.0078
22	'ал'	0.00774
23	'он'	0.0077
24	'ка'	0.00768
25	'ли'	0.00763
26	'ом'	0.00743
27	'ре'	0.00737
28	'пр'	0.00724
29	'ор'	0.00711
30	'ол'	0.00698
31	'ла'	0.00695
32	'ан'	0.00675
33	'ри'	0.00637
34	'ть'	0.00633
35	'ак'	0.0063
36	'те'	0.00625
37	'ве'	0.00622
38	'ло'	0.00615
39	'ел'	0.00611
40	'од'	0.00611

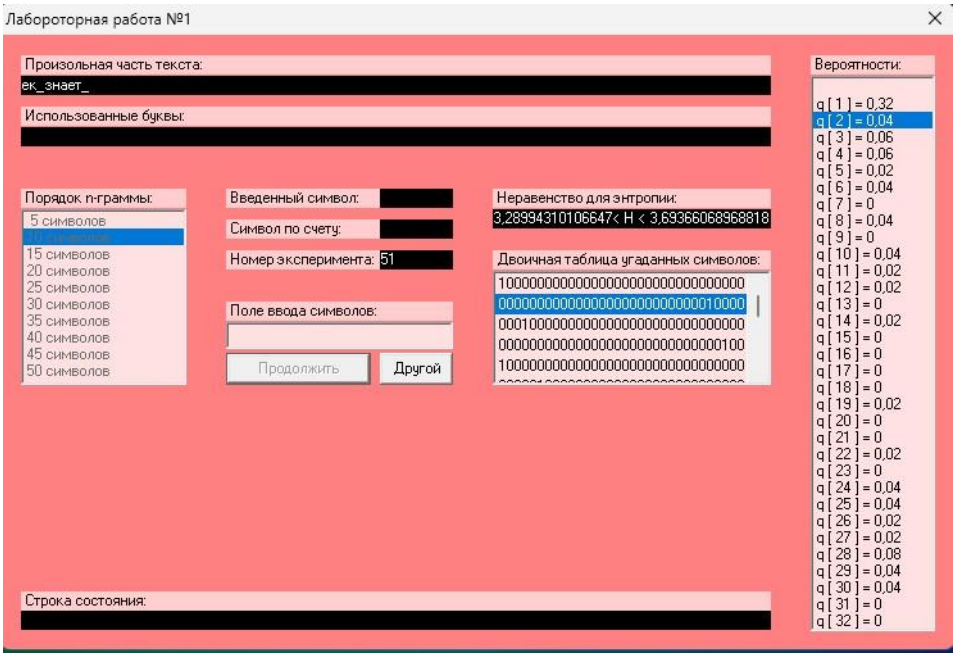
На першій таблиці частоти біграм з кроком 1 у тексті з пробілами, на другій частоти біграм з кроком 1 у тексті без пробілів, на третій та четвертій частоти біграм з кроком два у відповідному порядку стосовно пробілів.

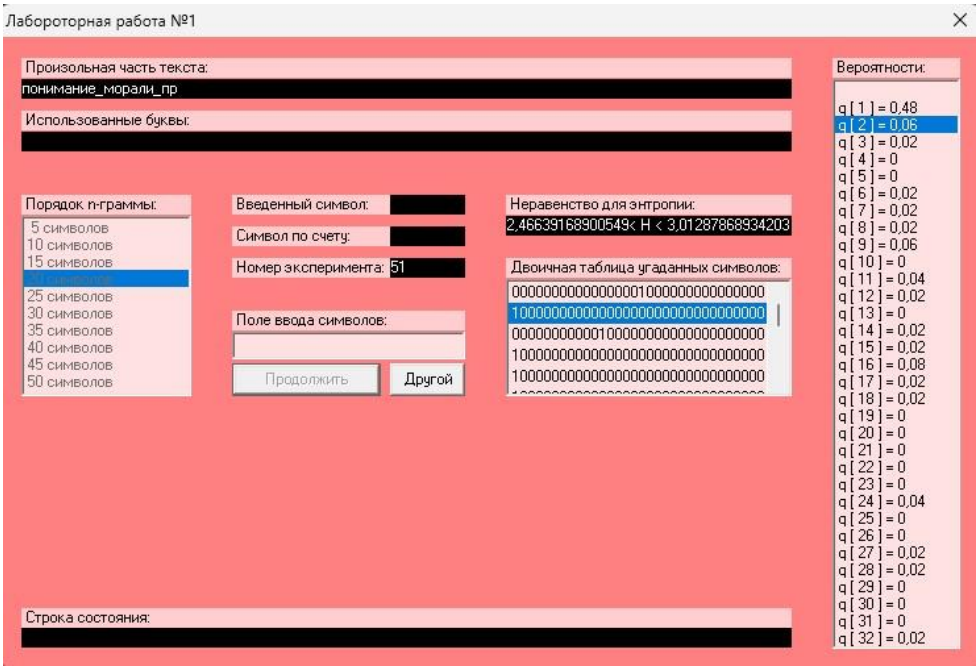
Одержані значення H та R:

ФБ-01 Чуйко Олександр, ФБ-01 Ченський Костянтин

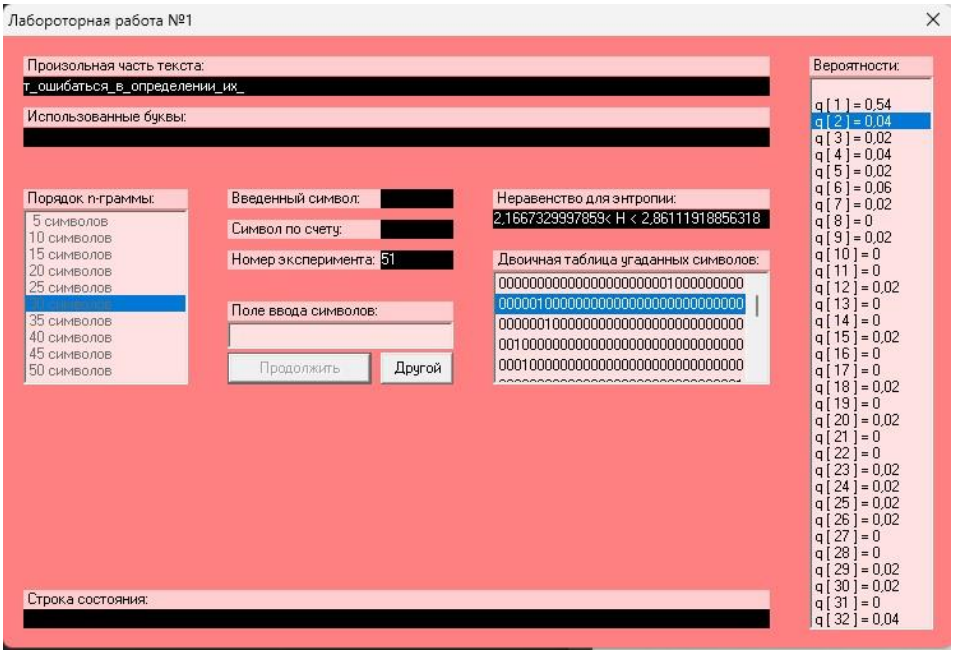
	A	B	C
1	**Name**	**H**	**R**
2	Letters with spaces	4.4115798996701	0.132852654195333
3	Letters with out spaces	4.48955901125098	0.109990435913449
4	Bigrams with spaces and step 1	4.03915965362009	0.206056185635473
5	Bigrams with out spaces and step 1	4.19671375043315	0.168044040348122
6	Bigrams with spaces and step 2	4.03887392578276	0.206112348765553
7	Bigrams with out spaces and step 2	4.19652114184497	0.16808222304829

Оцінки для $H^{(10)}$, $H^{(20)}$, $H^{(30)}$:





ФБ-01 Чуйко Олександр, ФБ-01 Ченський Костянтин



$3.28994310106647 < H^{(10)} < 3.69366068968818$

$0.342011379786706 < R^{(10)} < 0.261267862062364$

$2.46639168900549 < H^{(20)} < 3.01287868934203$

$0.506721662198902 < R^{(20)} < 0.397424262131594$

$2.16673299977859 < H^{(30)} < 2.86111918856318$

$0.566653400044282 < R^{(30)} < 0.427776162287364$

Висновки:

Виконавши цей комп’ютерний практикум ми набули практичних навичок щодо оцінки ентропії на символ джерела, порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії.

Написали програму для підрахунку усіх необхідних для роботи значень, використавши для збереження отриманих результатів модуль, який надає змогу запису даних у .xlsx файл.

Використавши отримані значення ентропії, підрахували надлишковість.

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