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

Лабораторна робота з криптографії

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Студенти групи ФБ-33

Бондар Марина Вікторівна,

Романовська Крістіна Миколаївна

Перевірив:

к.ф.-м.н., ст. викл.
кафедри математичних
методів захисту
інформації Селюх П.В

Київ 2025

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 частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

Для аналізу був обраний текст “Майстер і Маргарита” та записаний у файл Bulgakov_Mihail_Master_i_Margarita.txt
В подальшому програма буде зчитувати та аналізувати за допомогою функцій.

Усі отримані дані виконання скрипта були записані у таблицю “results.xlsx”

```
(venv) kristinaromanovska@MacBookAir lab1 % python3 lab1.py

Обробка

 З ПРОБІЛАМИ
H1 : 4.37282
H2 (крок 1): 3.98246
H2 (крок 2): 3.98177
R (H1): 0.12544
R (H2 крок 1): 0.20351
R (H2 крок 2): 0.20365

 БЕЗ ПРОБІЛІВ
H1 : 4.45071
H2 (крок 1): 4.14669
H2 (крок 2): 4.14651
R (H1): 0.10163
R (H2 крок 1): 0.16299
R (H2 крок 2): 0.16303
Результати збережено у файл.
```

Монограми

Монограми (з пробілами)

H1 : 4.37282
R (H1): 0.12544

Буква	Кількість	Частота
	113993	0,160041
о	66180	0,092913
а	51732	0,072629
е	48624	0,068265
и	40935	0,057470
н	38351	0,053843
т	36067	0,050636
л	31460	0,044168
с	30318	0,042565
р	28390	0,039858
в	28201	0,039592
к	21878	0,030715
у	17976	0,025237
м	17951	0,025202
п	17006	0,023875
д	16812	0,023603
г	11489	0,016130
я	11407	0,016014
ь	10820	0,015190
з	10632	0,014926

Монограми (без пробілів)

H1 : 4.45071
R (H1): 0.10163

Буква	Кількість	Частота
о	66180	0,110617
а	51732	0,086467
е	48624	0,081272
и	40935	0,068421
н	38351	0,064102
т	36067	0,060284
л	31460	0,052584
с	30318	0,050675
р	28390	0,047452
в	28201	0,047136
к	21878	0,036568
у	17976	0,030046
м	17951	0,030004
п	17006	0,028424
д	16812	0,028100
г	11489	0,019203
я	11407	0,019066
ь	10820	0,018085
з	10632	0,017770

Біграми, що перетинаються

Біграми крок 1 (з пробілами)

H2 (крок 1): 3.98246
R (H2 крок 1): 0.20351

	A	B	C
1	Біграма	Кількість	Частота
2	то	9705	0,016221
3	но	7071	0,011818
4	ст	6690	0,011182
5	на	6559	0,010963
6	по	6522	0,010901
7	ов	6474	0,010821
8	ал	6391	0,010682
9	не	6289	0,010511
10	ко	6135	0,010254
11	ро	5709	0,009542
12	ла	5539	0,009258
13	ен	5442	0,009096
14	го	5434	0,009082
15	ра	5388	0,009005
16	ос	5374	0,008982
17	от	5345	0,008933
18	он	5306	0,008868
19	ни	5188	0,008671
20	ка	5089	0,008506
21	ер	4901	0,008191

Біграми крок 1 (без пробілів)

H2 (крок 1): 4.14669
R (H2 крок 1): 0.1629

Біграма	Кількість	Частота
о	15225	0,021375
а	12853	0,018045
п	12090	0,016973
в	11619	0,016312
и	11555	0,016222
е	11551	0,016217
н	10906	0,015311
с	10513	0,014759
то	9417	0,013221
и	7885	0,011070
но	6913	0,009705
о	6851	0,009618
к	6851	0,009618
я	6693	0,009396
ь	6636	0,009316
ст	6573	0,009228
по	6520	0,009153

Біграми, що не перетинаються

Біграми крок 2 (з пробілами)

H2 (крок 2): 3.98177
R (H2 крок 2): 0.20365

Біграма	Кількість	Частота
о	7625	0,021410
а	6463	0,018147
п	5996	0,016836
и	5824	0,016353
е	5823	0,016350
в	5819	0,016339
н	5578	0,015662
с	5270	0,014797
то	4640	0,013028
и	3902	0,010956
но	3475	0,009757
о	3473	0,009751
к	3426	0,009619
ст	3395	0,009532
ь	3318	0,009316
я	3303	0,009274

Біграми крок 2 (без пробілів)

H2 (крок 2): 4.14651
R (H2 крок 2): 0.16303

	Біграма	Кількість	Частота
2	то	4794	0,016025
3	но	3593	0,012011
4	ст	3351	0,011202
5	ов	3298	0,011024
6	на	3253	0,010874
7	по	3206	0,010717
8	ал	3153	0,010540
9	не	3106	0,010383
10	ко	2973	0,009938
11	ро	2813	0,009403
12	ен	2761	0,009229
13	ла	2760	0,009226
14	ра	2760	0,009226
15	ос	2741	0,009162
16	го	2704	0,009039
17	от	2653	0,008868

Матриця біграм

Матриця біграм (з пробілами):

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
а	5,62E-06	0,000548	0,002507	0,000899	0,002234	0,001101	0,001085	0,003865	0,000293	0,000576	0,004246	0,008615	0,003097	0,005071	3,23E-05	0,000856
б	0,000952	1,12E-05	2,11E-05	1,83E-05	7,02E-06	0,002314	1,12E-05	1,4E-06	0,000658	0	0,000159	0,000567	4,49E-05	0,000316	0,001936	0
в	0,006328	1,12E-05	0,000111	1,97E-05	0,00027	0,005609	0	0,000413	0,003516	0	0,000232	0,000584	0,000117	0,000879	0,006291	0,000215
г	0,002135	0	2,81E-06	0	0,000899	0,000479	0	1,4E-06	0,000629	0	7,44E-05	0,001564	1,12E-05	0,000389	0,007597	0
д	0,004182	7,02E-05	0,000914	1,12E-05	1,12E-05	0,00385	5,19E-05	9,83E-06	0,00198	0	0,0003	0,00042	0,000129	0,001925	0,003252	0,000101
е	0,000131	0,000861	0,002081	0,002926	0,002805	0,00129	0,000758	0,001381	0,000199	0,001852	0,001412	0,005913	0,003642	0,006238	0,000194	0,001182
ж	0,001316	2,39E-05	0	1,4E-05	0,000716	0,002955	9,83E-06	0	0,001129	0	0,00014	4,07E-05	8,42E-06	0,000629	5,34E-05	0
з	0,005457	0,00017	0,001088	0,000272	0,000859	0,000628	8,7E-05	2,11E-05	0,00049	0	0,000112	0,000344	0,000234	0,001377	0,000647	0
и	0,000229	0,00056	0,002756	0,000489	0,001481	0,001651	0,000247	0,002189	0,000482	0,001617	0,00267	0,006127	0,002527	0,003228	0,000403	0,000296
й	0	9,83E-06	0	1,12E-05	9,97E-05	0	0	0	0	0	0,000121	1,26E-05	3,79E-05	0,000365	8,42E-06	0
к	0,007068	0	0,000539	0	0	0,000865	4,77E-05	1,54E-05	0,002881	0	2,25E-05	0,000806	2,81E-06	0,000674	0,008401	0
л	0,007506	7,86E-05	2,39E-05	0,000114	6,6E-05	0,004442	0,000212	2,95E-05	0,006225	0	0,000396	0,000427	5,34E-05	0,000459	0,006201	5,76E-05
м	0,003572	9,83E-06	6,88E-05	7,58E-05	0	0,003104	0	0	0,002799	0	0,000108	0,000181	8,42E-05	0,001262	0,002899	0,000171
н	0,009133	5,62E-06	2,95E-05	9,41E-05	0,000852	0,008787	1,12E-05	5,62E-05	0,007014	0	0,000258	0	0	0,003096	0,009706	0
о	8,42E-06	0,002972	0,006907	0,004324	0,004487	0,001526	0,00156	0,001474	0,000658	0,003806	0,002543	0,006229	0,00565	0,005284	0,000194	0,001341
п	0,001664	0	0	0	0	0,002121	0	0	0,001254	0	8,99E-05	0,000894	0	0,000254	0,009154	6,18E-05
р	0,007525	6,74E-05	0,000487	0,001011	0,000257	0,005598	0,000227	4,63E-05	0,005557	0	0,000396	0,000256	0,000195	0,000955	0,007911	6,32E-05
с	0,001664	7,02E-05	0,001476	2,67E-05	0,00025	0,002943	3,79E-05	1,83E-05	0,001797	0	0,003885	0,002464	0,000592	0,000861	0,002599	0,001608
т	0,005565	2,39E-05	0,001928	2,81E-06	0,00015	0,005697	0	4,21E-06	0,003424	0	0,000719	0,000212	2,81E-05	0,001245	0,013221	7,72E-05
у	0,000102	0,000635	0,000716	0,000928	0,001796	0,000174	0,001192	0,000278	2,11E-05	0,000146	0,000942	0,002183	0,000927	0,000496	4,21E-06	0,000615
ф	0,000272	0	0	0	0	0,000503	0	0	0,000334	0	0	8,84E-05	0	1,4E-06	0,000244	0
х	0,000567	1,4E-06	0,000166	5,76E-05	0	6,32E-05	0	0	0,000313	0	7,02E-06	0,000163	5,76E-05	0,000327	0,001776	0
ц	0,00063	0	6,32E-05	0	0	0,00083	0	0	0,000284	0	1,4E-05	5,62E-06	0	0	0,000257	0
ч	0,002242	0	1,4E-06	0	0	0,003782	0	0	0,001547	0	0,00049	2,53E-05	4,21E-06	0,000776	6,32E-05	0
ш	0,001018	0	6,46E-05	0	0	0,002054	0	0	0,001801	0	0,00039	0,000428	7,02E-06	0,00038	0,000241	4,77E-05
щ	0,000441	0	0	0	0	0,001244	0	0	0,001026	0	0	0	0	5,48E-05	4,21E-06	0

Матриця біграм (без пробілів):

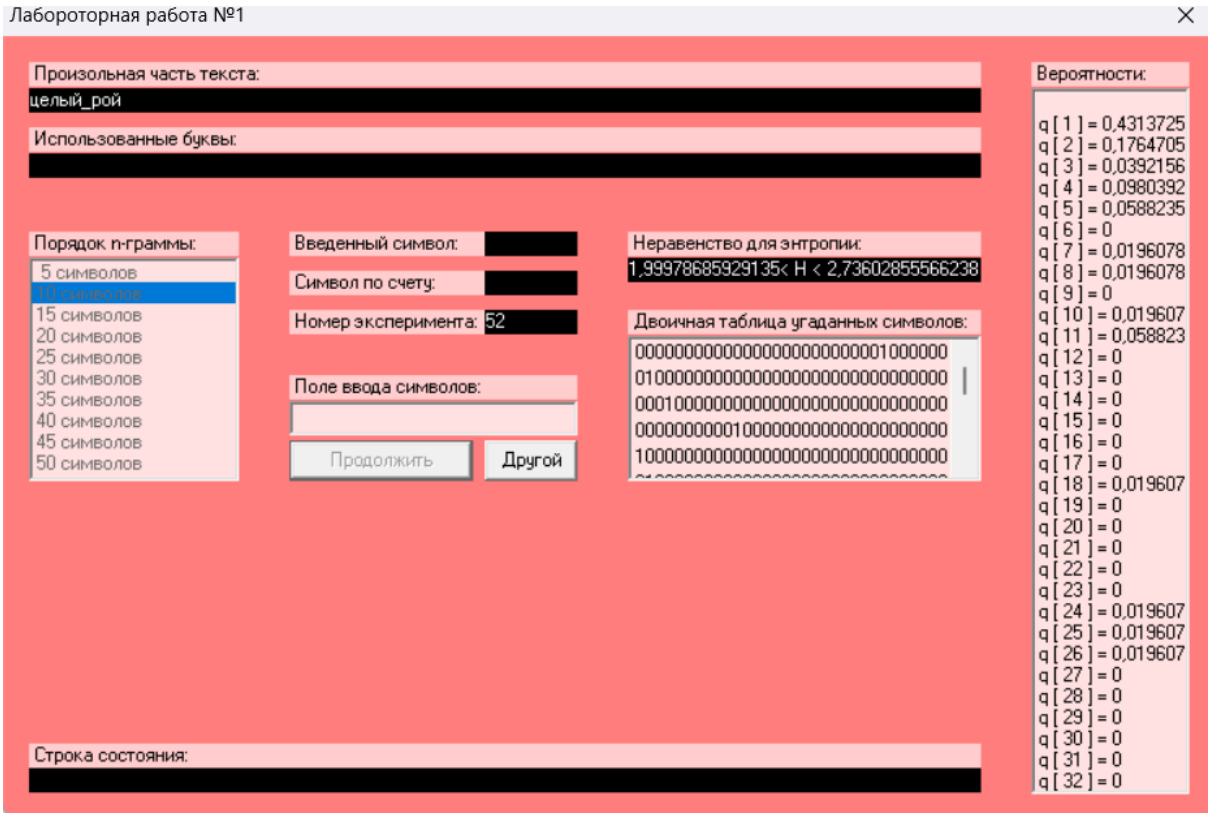
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
а	0,000471	0,001422	0,005121	0,001707	0,003396	0,001747	0,001494	0,00521	0,001844	0,000687	0,006373	0,010682	0,004588	0,008172	0,001193	0,003435
б	0,001137	1,5E-05	4,35E-05	2,34E-05	1E-05	0,002758	1,5E-05	5,01E-06	0,000812	0	0,000197	0,000675	6,18E-05	0,000383	0,002337	1,84E-05
в	0,007615	0,000276	0,000623	0,000425	0,000605	0,006808	7,19E-05	0,000649	0,004451	3,34E-06	0,001135	0,000929	0,00047	0,00159	0,007966	0,0011
г	0,002557	3,01E-05	5,68E-05	1E-05	0,001096	0,00058	3,34E-06	4,51E-05	0,000812	0	0,000114	0,001867	2,67E-05	0,000506	0,009083	7,86E-05
д	0,005031	0,000127	0,001225	8,36E-05	6,02E-05	0,004618	6,52E-05	8,52E-05	0,002469	0	0,000485	0,000545	0,000234	0,002517	0,003933	0,000301
е	0,000446	0,001986	0,004429	0,003953	0,004	0,001792	0,001125	0,002345	0,001468	0,00221	0,002571	0,007426	0,00518	0,009096	0,001362	0,003619
ж	0,001566	4,68E-05	2,67E-05	2,17E-05	0,000866	0,003535	1,17E-05	1E-05	0,001366	0	0,000192	5,18E-05	1,84E-05	0,000796	8,36E-05	1,34E-05
з	0,00654	0,000261	0,001476	0,000389	0,001115	0,000766	0,000115	0,000152	0,000657	1,67E-06	0,000331	0,00044	0,000383	0,00184	0,000876	0,000242
и	0,000525	0,001404	0,005579	0,001157	0,00246	0,002283	0,000443	0,003279	0,001681	0,001926	0,004227	0,007637	0,003642	0,005653	0,001732	0,002576
й	0,000197	0,000394	0,000816	0,00043	0,000511	8,02E-05	0,000157	0,000236	0,00069	0	0,000839	0,000184	0,000478	0,001105	0,00046	0,001061
к	0,008506	0,000374	0,001172	0,000115	0,000125	0,001098	0,000181	0,000119	0,003828	0	0,000309	0,001036	0,000192	0,00127	0,010254	0,000438
л	0,009258	0,000495	0,00115	0,000425	0,000321	0,005634	0,000289	0,000219	0,008182	0	0,001148	0,000605	0,000374	0,001417	0,007889	0,001028
м	0,004376	0,000358	0,000928	0,000468	0,000323	0,003818	9,86E-05	0,000254	0,003983	0	0,000607	0,000418	0,000356	0,002282	0,004043	0,001117
н	0,010963	0,000192	0,000428	0,000194	0,001122	0,010512	5,18E-05	0,000209	0,008672	0	0,000478	3,84E-05	0,000107	0,00414	0,011819	0,000503
о	0,000303	0,004719	0,010821	0,005711	0,006196	0,002333	0,002285	0,002492	0,002216	0,004531	0,004249	0,007824	0,00762	0,008869	0,002078	0,004117
п	0,001982	3,34E-06	1,34E-05	1,67E-06	1,67E-06	0,002526	0	3,34E-06	0,001501	0	0,000109	0,001065	1,67E-06	0,000306	0,010901	7,52E-05
р	0,009006	0,000125	0,000747	0,001234	0,000364	0,006684	0,000277	9,03E-05	0,006947	0	0,000565	0,000329	0,000271	0,001322	0,009542	0,000284
с	0,002034	0,000249	0,002056	0,00017	0,000403	0,003555	0,000147	0,00012	0,002295	0	0,004943	0,002994	0,000824	0,001422	0,003251	0,00229
т	0,006776	0,000271	0,002975	0,00017	0,000373	0,00694	0,00017	0,000192	0,004583	0	0,001269	0,000398	0,000267	0,002138	0,016222	0,000779
у	0,000314	0,000991	0,001666	0,001244	0,002412	0,000304	0,001494	0,000506	0,000847	0,000174	0,001559	0,002701	0,001437	0,001524	0,000394	0,001534
ф	0,000324	1,67E-06	5,01E-06	0	1,67E-06	0,000598	1,67E-06	0	0,000401	0	1,67E-06	0,000105	0	8,36E-06	0,000299	3,34E-06
х	0,000752	0,00013	0,000485	0,00016	0,00015	0,000115	2,67E-05	0,000105	0,000654	0	0,000286	0,000304	0,000209	0,00066	0,0023	0,000443
ц	0,000757	1,67E-05	0,000115	1E-05	1,67E-05	0,000993	0	8,36E-06	0,000383	0	3,68E-05	1,17E-05	1E-05	3,68E-05	0,000341	4,51E-05
ч	0,002689	3,68E-05	6,35E-05	1,5E-05	1,34E-05	0,004505	1,67E-06	4,18E-05	0,001872	0	0,00061	4,68E-05	6,69E-06	0,000996	0,000124	8,02E-05
ш	0,001215	0	8,36E-05	3,34E-06	6,69E-06	0,002445	0	1,67E-06	0,002153	0	0,000465	0,000515	8,36E-06	0,000465	0,000291	7,19E-05
щ	0,000528	5,01E-06	5,01E-06	0	1,67E-06	0,001483	0	0	0,001224	0	3,34E-06	1,67E-06	0	6,52E-05	8,36E-06	5,01E-06

2. За допомогою програми CoolPinkProgram оцінити значення H^{10} , H^{20} , H^{30}.

Отже, після виконання завдання ми отримали результати для кожної ентропії n-грам.

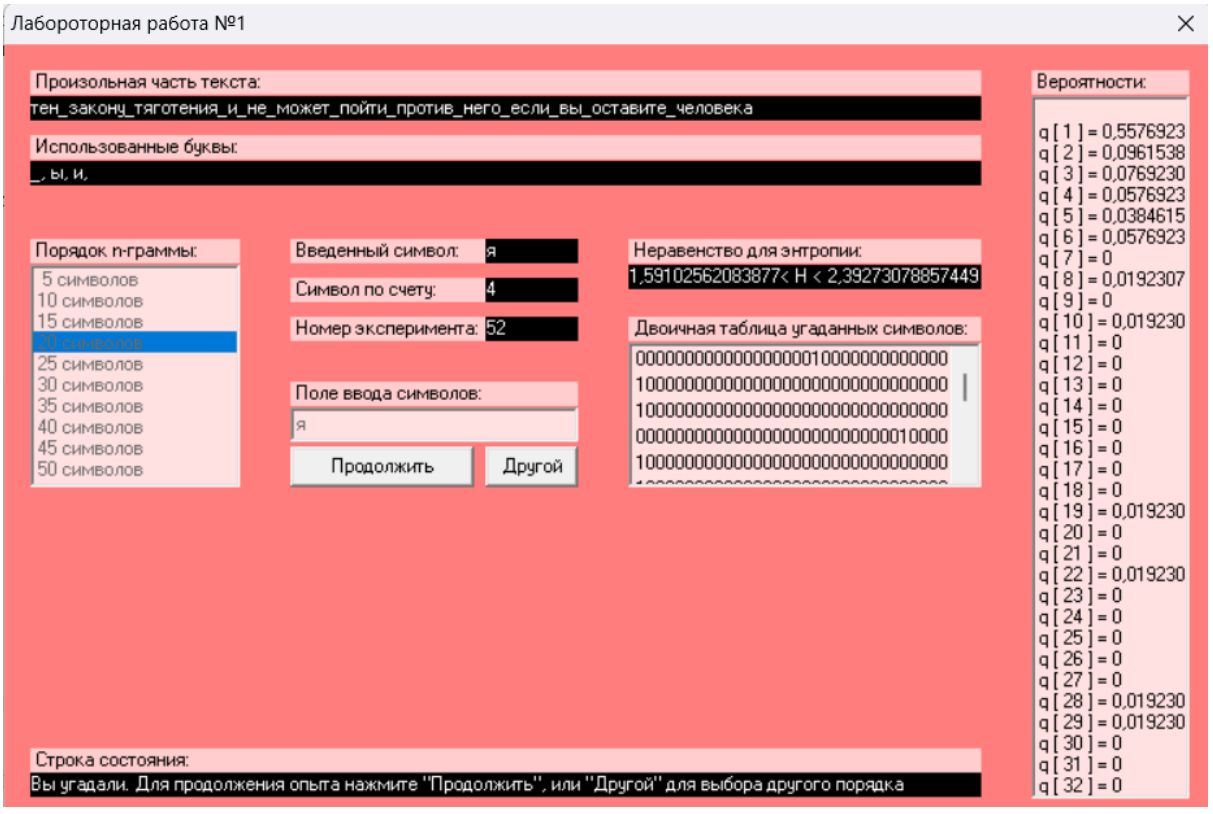
1) **H^{10}**

1,99978685929135< H^{10} < 2,73602285566283



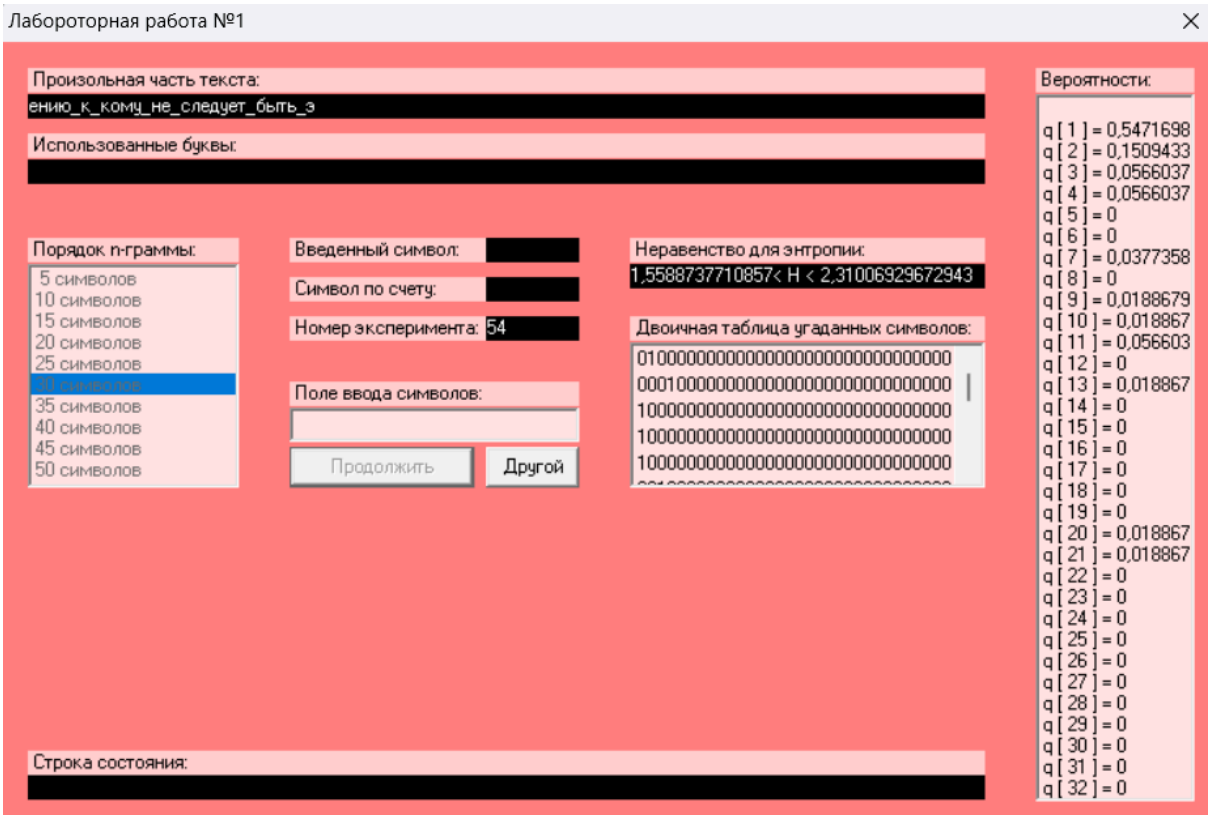
2) H^{20}

$1{,}59103562083877 < H^{20} < 2{,}39273078857449$



3) H^{30}

1.5588737710857 < H^{30} < 2.31006929672943



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

Для виконання цього пункту, ми використали формули із матеріалів комп'ютерного практикуму:

Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \frac{H_\infty}{H_0}$

$H_0 = \log_2 m$, де m – кількість букв в алфавіті Z.

Оскільки в рос. алфавіті є 33 літери, за умовами здачі "Ё" та "Ъ" замінюємо на "Е" та "Ь", пробіл вважаємо окремою літерою. У результаті ми отримали алфавіт, яким налічує 32 букви.

$H_0 = \log_2 32 = 5$

• H10

$R_l = 1 - \frac{1,99978685929135}{5} = 1 - 0.39995737185827 = 0.60$

$R_r = 1 - \frac{2,73602285566283}{5} = 1 - 0.54720457113257 = 0.45$

0.60 < H10 < 0.45

• H20

$R_l = 1 - \frac{1.59103562083877}{5} = 1 - 0.31820712416775 = 0.68179287583225 = 0.68$

$R_r = 1 - \frac{2.39273078857419}{5} = 1 - 0.47854615771490 = 0.52145384228510 = 0.52$

$0.68 < H120 < 0.52$

• H30

$R_l = 1 - \frac{1.5588737710857}{5} = 1 - 0.31177475421714 = 0.68822524578286 = 0.69$

$R_r = 1 - \frac{2.31006929672943}{5} = 1 - 0.46201385934589 = 0.53798614065411 = 0.54$

$0.69 < H3° < 0.54$

Висновки: у ході виконання комп'ютерного практикуму, було досліджено ентропію та надлишковість рос. мови. За допомогою статичного аналізу скриптом, ми змогли застосувати навички визначення частот могонрами, біграми. На основі цих роздахунків було отримано ентропію першого та другого порядку (монограм та біграм).

Експериментальна оцінка, яка була здійснена за допомогою програми CoolPinkProgram, продемонструвала, що метод вгадування наступних літер, зменшує ентропію із збільшенням порядку (H^{10} , H^{20} , H^{30})