Комп’ютерний практикум №1

Експерментальна оцінка ентропії на символ джерела відкритого тексту

Кузьменко Вікторія та Будніков Дмитро

Результати

Частота символів без пробілів

літера	частота
о	0.11472849174209941
е	0.08663889219734265
а	0.0757714487057329
н	0.06735639656450161
и	0.06312777910115928
т	0.06207279590810439
с	0.05702740392250278
л	0.049138395929278496
в	0.042681534857868825
р	0.04241065520618284
м	0.03597943465142131
к	0.034031168942882856
д	0.031233044174474617
у	0.028277767852416494
п	0.026429582870149806
я	0.021581043883330688
ь	0.019097222222222224
ы	0.01903560227092266
г	0.017507096633317454
з	0.016561705971097348
б	0.016421510242972844
ч	0.01558281721454661
й	0.011790915647662908
ж	0.010193346037795776
х	0.009091217378645915
ш	0.008231846514213118
ю	0.005641327219310783
э	0.004571042427611032
ц	0.0031558513313217932
щ	0.003054943491609761
ф	0.001577718887300831

Частота символів з пробілами

літера	частота
	0.164822958416055
о	0.09581860231855466
е	0.07235881367148697
а	0.06328257436658363
н	0.05625451601449546
и	0.052722871791471004
т	0.05184177404937463
с	0.04762797849720853
л	0.04103926014039538
в	0.03564663801285691
р	0.03542040554673652
м	0.030049197790036928
к	0.02842205099936033
д	0.02608512143329838
у	0.023616942497578796
п	0.022073380831789426
я	0.018023992384773414
ь	0.015949561558026727
ы	0.015898097989397814
г	0.014621525172938315
з	0.013831956596524241
б	0.01371486834306651
ч	0.013014411180788409
й	0.009847502048180953
ж	0.008513248587687703
х	0.0075927760346940435
ш	0.0068750492185136215
ю	0.004711506977630963
э	0.0038176296916468755
ц	0.0026356945785720896
щ	0.0025514186675287677
ф	0.0013176745927470216

Частота перехресних біграм без пробілів

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н
а	0.000219	0.001470	0.004650	0.001340	0.003160	0.001850	0.001400	0.004520	0.001160	0.001080	0.006420	0.008640	0.005240	0.0
б	0.000866	0.000051	0.000079	0.000009	0.000028	0.002120	0.000014	0.000009	0.000986	0.000000	0.000158	0.000856	0.000060	0.0
в	0.006640	0.000207	0.000348	0.000319	0.000715	0.006530	0.000046	0.000628	0.003810	0.000002	0.000916	0.001080	0.000402	0.0
г	0.001390	0.000073	0.000086	0.000020	0.001490	0.000464	0.000005	0.000031	0.000838	0.000000	0.000157	0.001630	0.000058	0.0
д	0.005010	0.000073	0.001270	0.000045	0.000113	0.005960	0.000540	0.000058	0.002740	0.000000	0.000424	0.000784	0.000111	0.0
е	0.000226	0.001880	0.003460	0.004560	0.004750	0.002020	0.001160	0.002240	0.001170	0.002640	0.002500	0.007130	0.005840	0.0
ж	0.001400	0.000040	0.000024	0.000016	0.000959	0.004130	0.000012	0.000014	0.001580	0.000000	0.000157	0.000013	0.000024	0.0
з	0.006160	0.000212	0.001090	0.000457	0.000992	0.000428	0.000108	0.000096	0.000447	0.000000	0.000311	0.000379	0.000476	0.0
и	0.000349	0.001450	0.004220	0.000832	0.002810	0.002930	0.000427	0.002960	0.001670	0.001300	0.003210	0.005770	0.004140	0.0
й	0.000139	0.000325	0.000740	0.000199	0.000941	0.000108	0.000169	0.000222	0.000514	0.000000	0.000799	0.000234	0.000586	0.0
к	0.008370	0.000246	0.000802	0.000069	0.000239	0.000938	0.000127	0.000151	0.003060	0.000000	0.000364	0.001040	0.000273	0.0
л	0.005320	0.000338	0.000868	0.000275	0.000366	0.005550	0.000363	0.000203	0.008340	0.000000	0.000846	0.000450	0.001840	0.0
м	0.002770	0.000391	0.000902	0.000232	0.000558	0.005080	0.000100	0.000246	0.004230	0.000000	0.000595	0.000344	0.000556	0.0
н	0.012600	0.000332	0.000621	0.000442	0.000840	0.011800	0.000044	0.000178	0.008800	0.000000	0.000646	0.000062	0.000188	0.0
о	0.000199	0.005460	0.009680	0.005860	0.006250	0.003420	0.002800	0.002250	0.002320	0.004580	0.004200	0.008320	0.008160	0.0
п	0.001500	0.000006	0.000006	0.000002	0.000007	0.002220	0.000000	0.000001	0.001310	0.000000	0.000133	0.000847	0.000006	0.0
р	0.008160	0.000156	0.000669	0.000287	0.000560	0.006680	0.000373	0.000084	0.005280	0.000000	0.000447	0.000429	0.000416	0.0
с	0.001920	0.000267	0.002260	0.000133	0.000517	0.003950	0.000109	0.000116	0.002000	0.000000	0.004780	0.003990	0.001250	0.0
т	0.006460	0.000315	0.003170	0.000113	0.000407	0.007440	0.000092	0.000158	0.004600	0.000000	0.001050	0.000521	0.000402	0.0
у	0.000198	0.001010	0.001650	0.001400	0.002560	0.000404	0.001670	0.000557	0.000622	0.000166	0.001290	0.001730	0.001860	0.0

Частота перехресних біграм з пробілами

	а	б	в	г	д	е	ж	з	и	й	к	л	м
а	0.000002	0.000627	0.002472	0.000834	0.001986	0.001177	0.001058	0.003378	0.000118	0.000899	0.004482	0.006952	0.0
б	0.000713	0.000040	0.000050	0.000004	0.000017	0.001767	0.000009	0.000003	0.000793	0.000000	0.000122	0.000715	0.0
в	0.005445	0.000008	0.000029	0.000012	0.000273	0.005319	0.000000	0.000429	0.003026	0.000000	0.000152	0.000726	0.0
г	0.001149	0.000004	0.000005	0.000007	0.001197	0.000372	0.000000	0.000001	0.000662	0.000000	0.000094	0.001349	0.0
д	0.004160	0.000028	0.000963	0.000009	0.000046	0.004945	0.000446	0.000009	0.002201	0.000000	0.000286	0.000612	0.0
е	0.000036	0.000762	0.001175	0.003485	0.003179	0.001337	0.000764	0.001268	0.000091	0.002203	0.001283	0.005630	0.0
ж	0.001167	0.000025	0.000001	0.000010	0.000791	0.003441	0.000010	0.000000	0.001302	0.000000	0.000117	0.000008	0.0
з	0.005124	0.000141	0.000789	0.000340	0.000729	0.000332	0.000079	0.000009	0.000331	0.000000	0.000140	0.000275	0.0
и	0.000133	0.000531	0.001734	0.000424	0.001614	0.002023	0.000256	0.002042	0.000629	0.001088	0.001904	0.004618	0.0
й	0.000003	0.000016	0.000018	0.000001	0.000307	0.000002	0.000000	0.000013	0.000001	0.000000	0.000165	0.000052	0.0
к	0.006952	0.000002	0.000169	0.000000	0.000003	0.000722	0.000029	0.000036	0.002290	0.000000	0.000034	0.000794	0.0
л	0.004395	0.000033	0.000048	0.000115	0.000099	0.004288	0.000280	0.000017	0.006632	0.000000	0.000352	0.000285	0.0
м	0.002211	0.000060	0.000001	0.000040	0.000001	0.004094	0.000000	0.000010	0.003025	0.000000	0.000078	0.000148	0.0
н	0.010464	0.000020	0.000039	0.000297	0.000526	0.009801	0.000011	0.000021	0.007063	0.000000	0.000340	0.000011	0.0
о	0.000008	0.003572	0.005651	0.004593	0.004201	0.002349	0.002021	0.001243	0.000933	0.003826	0.002559	0.006591	0.0
п	0.001249	0.000000	0.000001	0.000000	0.000000	0.001852	0.000000	0.000000	0.001084	0.000000	0.000104	0.000707	0.0
р	0.006786	0.000063	0.000425	0.000163	0.000383	0.005551	0.000306	0.000040	0.004311	0.000000	0.000284	0.000334	0.0
с	0.001561	0.000061	0.001428	0.000033	0.000224	0.003200	0.000023	0.000007	0.001510	0.000000	0.003725	0.003264	0.0
т	0.005325	0.000025	0.002062	0.000009	0.000083	0.006070	0.000000	0.000005	0.003548	0.000000	0.000581	0.000335	0.0
у	0.000054	0.000687	0.000672	0.001077	0.001877	0.000235	0.001308	0.000310	0.000070	0.000138	0.000734	0.001368	0.0
ф	0.000209	0.000000	0.000000	0.000007	0.000000	0.000253	0.000000	0.000000	0.000200	0.000000	0.000000	0.000036	0.0
х	0.000547	0.000000	0.000169	0.000001	0.000000	0.000066	0.000000	0.000000	0.000209	0.000000	0.000001	0.000089	0.0
ц	0.000557	0.000000	0.000047	0.000000	0.000000	0.000697	0.000000	0.000000	0.000382	0.000000	0.000036	0.000000	0.0

Частота неперехресних біграм з пробілами

	а	б	в	г	д	е	ж	з	и	й	к	л	м
а	0.000001	0.000633	0.002454	0.000822	0.001966	0.001221	0.001064	0.003372	0.000125	0.000912	0.004491	0.006906	0.0034
б	0.000685	0.000040	0.000052	0.000004	0.000016	0.001778	0.000010	0.000004	0.000796	0.000000	0.000117	0.000707	0.0000
в	0.005443	0.000010	0.000036	0.000014	0.000274	0.005323	0.000001	0.000448	0.003004	0.000000	0.000146	0.000747	0.0001
г	0.001158	0.000002	0.000008	0.000006	0.001184	0.000388	0.000000	0.000002	0.000651	0.000000	0.000090	0.001386	0.0000
д	0.004195	0.000031	0.000949	0.000008	0.000047	0.004963	0.000457	0.000010	0.002251	0.000000	0.000280	0.000608	0.0000
е	0.000035	0.000782	0.001143	0.003494	0.003125	0.001320	0.000770	0.001255	0.000091	0.002150	0.001277	0.005614	0.0037
ж	0.001169	0.000025	0.000001	0.000012	0.000801	0.003414	0.000010	0.000000	0.001266	0.000000	0.000114	0.000010	0.0000
з	0.005082	0.000143	0.000780	0.000345	0.000725	0.000336	0.000088	0.000011	0.000333	0.000000	0.000144	0.000283	0.0003
и	0.000128	0.000553	0.001741	0.000403	0.001643	0.002001	0.000243	0.002093	0.000597	0.001070	0.001907	0.004640	0.0026
й	0.000001	0.000012	0.000017	0.000001	0.000307	0.000002	0.000001	0.000012	0.000001	0.000000	0.000158	0.000047	0.0001
к	0.006952	0.000003	0.000171	0.000000	0.000004	0.000734	0.000032	0.000035	0.002244	0.000000	0.000030	0.000789	0.0000
л	0.004433	0.000032	0.000052	0.000116	0.000107	0.004276	0.000271	0.000018	0.006649	0.000000	0.000336	0.000272	0.0012
м	0.002248	0.000061	0.000001	0.000040	0.000001	0.004134	0.000000	0.000009	0.003007	0.000000	0.000083	0.000141	0.0000
н	0.010428	0.000018	0.000044	0.000298	0.000508	0.009862	0.000012	0.000014	0.007067	0.000000	0.000332	0.000010	0.0000
о	0.000006	0.003570	0.005682	0.004503	0.004221	0.002343	0.002020	0.001257	0.000944	0.003841	0.002548	0.006512	0.0055
п	0.001231	0.000000	0.000001	0.000001	0.000000	0.001907	0.000000	0.000000	0.001103	0.000000	0.000099	0.000699	0.0000
р	0.006877	0.000060	0.000414	0.000163	0.000394	0.005626	0.000293	0.000041	0.004275	0.000000	0.000296	0.000331	0.0002
с	0.001532	0.000053	0.001453	0.000037	0.000216	0.003173	0.000027	0.000006	0.001518	0.000000	0.003738	0.003232	0.0008
т	0.005328	0.000025	0.002059	0.000008	0.000076	0.005957	0.000000	0.000004	0.003541	0.000000	0.000600	0.000331	0.0000
у	0.000054	0.000694	0.000664	0.001082	0.001898	0.000244	0.001337	0.000332	0.000065	0.000139	0.000759	0.001387	0.0010
ф	0.000213	0.000000	0.000000	0.000006	0.000000	0.000256	0.000000	0.000000	0.000216	0.000000	0.000000	0.000031	0.0000

Частота неперехресний біграм без пробілів

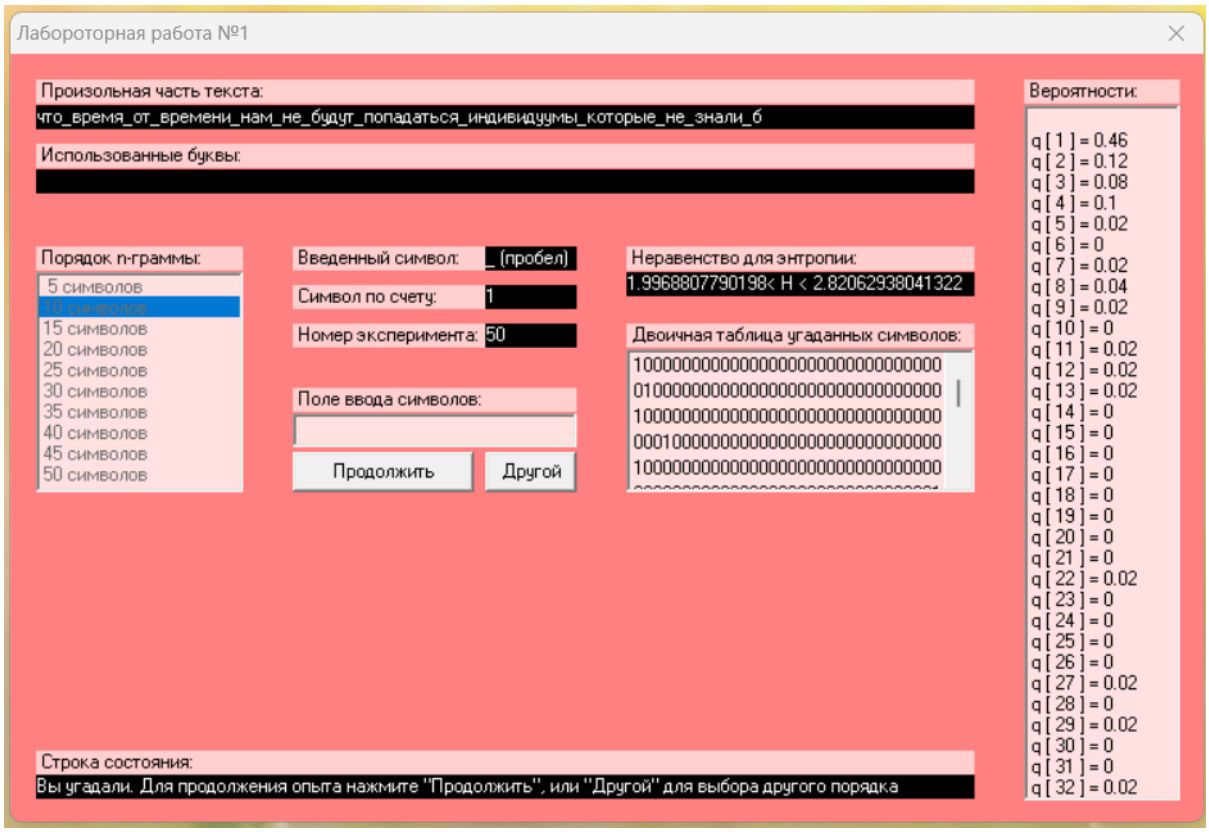
	а	б	в	г	д	е	ж	з	и	й	к	л	м	н
а	0.000212	0.001473	0.004626	0.001356	0.003070	0.001849	0.001383	0.004509	0.001167	0.001079	0.006383	0.008669	0.005258	0.0
б	0.000884	0.000045	0.000077	0.000007	0.000031	0.002184	0.000011	0.000005	0.000966	0.000000	0.000148	0.000874	0.000052	0.0
в	0.006566	0.000198	0.000342	0.000326	0.000684	0.006479	0.000050	0.000634	0.003724	0.000002	0.000929	0.001070	0.000371	0.0
г	0.001378	0.000076	0.000083	0.000020	0.001512	0.000469	0.000005	0.000025	0.000839	0.000000	0.000139	0.001600	0.000055	0.0
д	0.005050	0.000071	0.001305	0.000050	0.000098	0.005878	0.000544	0.000060	0.002720	0.000000	0.000414	0.000754	0.000110	0.0
е	0.000223	0.001859	0.003459	0.004622	0.004808	0.001963	0.001143	0.002260	0.001163	0.002629	0.002400	0.007161	0.005839	0.0
ж	0.001412	0.000035	0.000021	0.000016	0.000967	0.004167	0.000015	0.000008	0.001562	0.000000	0.000159	0.000012	0.000022	0.0
з	0.006171	0.000204	0.001100	0.000468	0.001016	0.000424	0.000112	0.000103	0.000445	0.000000	0.000309	0.000381	0.000472	0.0
и	0.000352	0.001458	0.004308	0.000862	0.002811	0.002935	0.000422	0.002844	0.001653	0.001304	0.003160	0.005752	0.004136	0.0
й	0.000137	0.000334	0.000732	0.000202	0.000955	0.000108	0.000168	0.000240	0.000523	0.000001	0.000807	0.000238	0.000576	0.0
к	0.008432	0.000258	0.000813	0.000076	0.000253	0.000931	0.000124	0.000142	0.003131	0.000000	0.000350	0.001055	0.000285	0.0
л	0.005352	0.000322	0.000862	0.000285	0.000368	0.005479	0.000352	0.000206	0.008352	0.000000	0.000843	0.000452	0.001850	0.0
м	0.002823	0.000408	0.000907	0.000222	0.000567	0.005189	0.000101	0.000242	0.004198	0.000001	0.000599	0.000356	0.000577	0.0
н	0.012680	0.000332	0.000608	0.000462	0.000830	0.011875	0.000047	0.000171	0.008908	0.000000	0.000656	0.000060	0.000176	0.0
о	0.000197	0.005388	0.009748	0.005861	0.006372	0.003442	0.002837	0.002217	0.002312	0.004505	0.004278	0.008393	0.008122	0.0
п	0.001501	0.000008	0.000007	0.000003	0.000007	0.002136	0.000000	0.000000	0.001325	0.000000	0.000135	0.000849	0.000007	0.0
р	0.008142	0.000170	0.000661	0.000285	0.000576	0.006581	0.000364	0.000081	0.005339	0.000000	0.000434	0.000423	0.000404	0.0
с	0.001965	0.000266	0.002209	0.000122	0.000519	0.003930	0.000108	0.000124	0.001984	0.000000	0.004738	0.004021	0.001217	0.0
т	0.006459	0.000323	0.003202	0.000110	0.000403	0.007360	0.000098	0.000149	0.004676	0.000000	0.001017	0.000500	0.000394	0.0

Та значення ентропії та надлишковості:

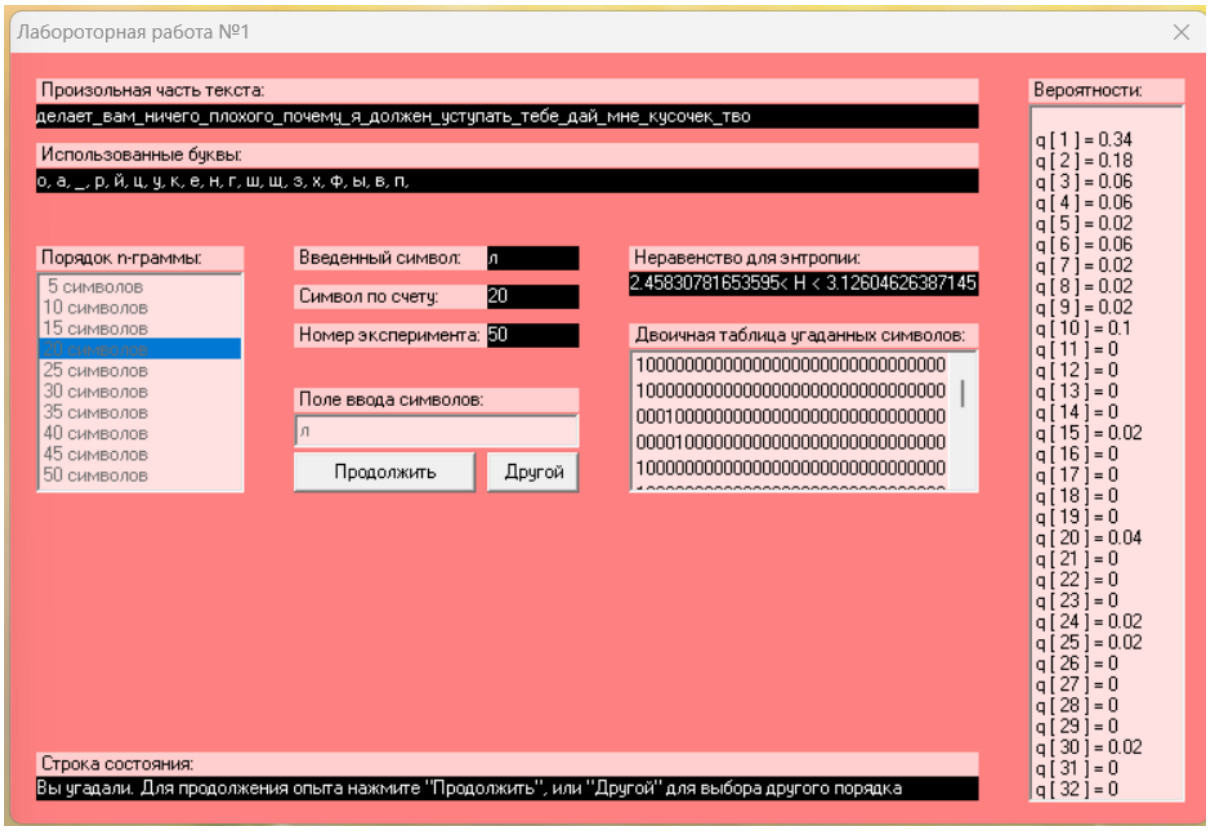
```
Ентропія літер (з пробілами): 4.36932                    R = 0.12614
Ентропія літер (без пробілів): 4.45845                   R = 0.10007

Ентропія перехресних біграм (з пробілами): 3.97790       R = 0.20442
Ентропія перехресних біграм (без пробілів): 4.15282      R = 0.16176

Ентропія неперехресних біграм (з пробілами): 3.97756     R = 0.20449
Ентропія неперехресних біграм (без пробілів): 4.15258    R = 0.16180
```

Перейдемо до роботи з програмою CoolPinkProgram.exe:

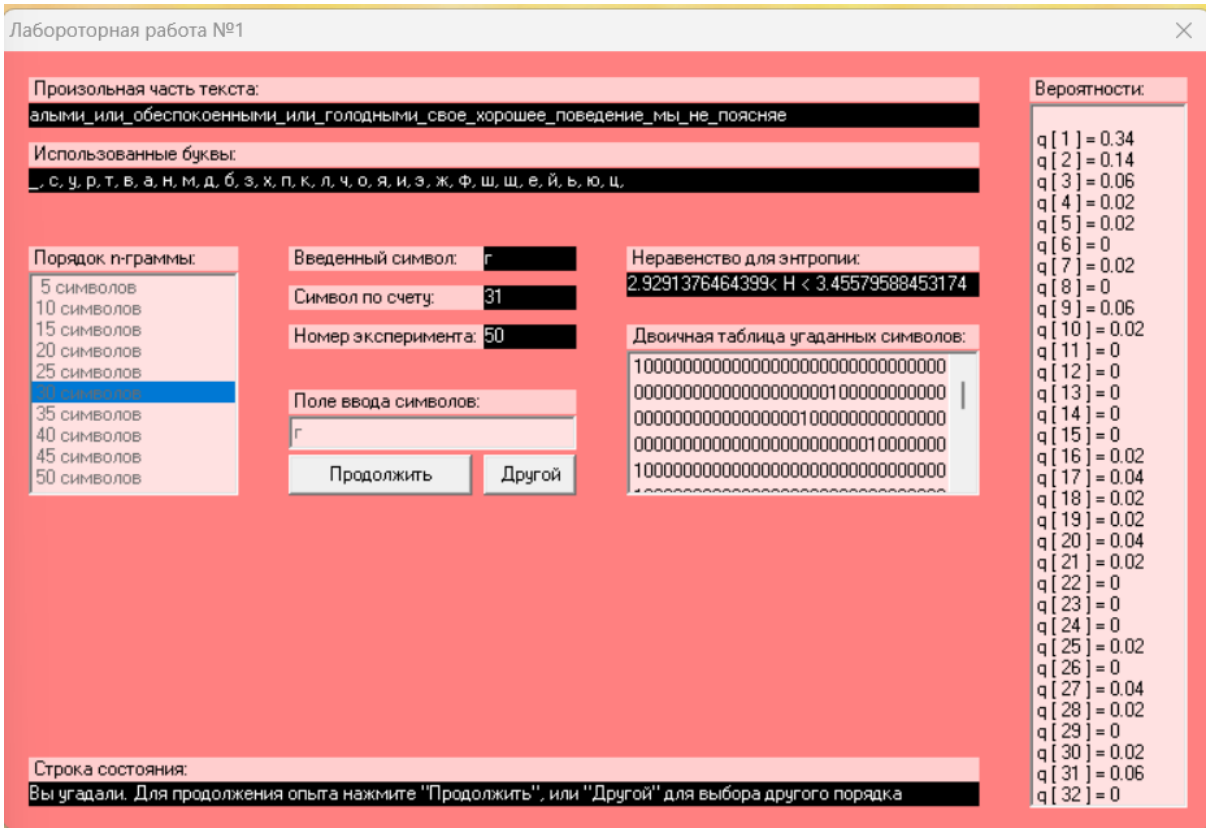
1. $H^{(10)}$



2. $H^{(20)}$



3. $H^{(30)}$



$H^{(x)}$	Нижня межа	Верхня межа
$H^{(10)}$	1.9968807790198	2.82062938041322
$H^{(20)}$	2.45830781653595	3.12604626387145
$H^{(30)}$	2.9291376464399	3.45579588453174

Було помічено зростання ентропії. Чим більший порядок n-грами ми обирали, тим вищими ставали як нижня, так і верхня межі ентропії. Це свідчить про збільшення кількості інформації та складності тексту при аналізі довших послідовностей символів.

Здійснимо оцінку надлишковості рос. Мови в різних моделях джерала згідно з формули:

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

1) $H^{(10)}$

H_0 - максимальна етропія алфавіту, рівна $\log_2(m)$, де $m = 32$ - кількість символів в алфавіті $\Rightarrow \log_2 32 = 5$

$$R_1 = 1 - \frac{1.9968807790198}{5} = 0.6006238442$$

$$R_2 = 1 - \frac{2.82062938041322}{5} = 0.4358741239$$

2) $H^{(20)}$

$$R_1 = 1 - \frac{2.45830781653595}{5} = 0.5083384367$$

$$R_2 = 1 - \frac{3.12604626387145}{5} = 0.3747907472$$

3) $H^{(30)}$

$$R_1 = 1 - \frac{2.9291376464399}{5} = 0.4141724707$$

$$R_2 = 1 - \frac{3.45579588453174}{5} = 0.3088408231$$

$R^{(x)}$	Нижня межа	Верхня межа
$R^{(10)}$	0.6006238442	0.4358741239
$R^{(20)}$	0.5083384367	0.3747907472
$R^{(30)}$	0.4141724707	0.3088408231

Висновки:

У результаті було досліджено ентропію російського мови та її надлишковість. Після здійснення оцінки надлишковості, було зрозуміло, що з тексту можна відкинути за певним алгоритмом 60% його символів, таким чином, що його зміст не буде втрачено. Зрозуміло, також те, що чим більші nграми буде розглянуто, тим більша буде точність надлишковості