



Комп'ютерний практикум №1
Експериментальна оцінка ентропії на символ
джерела відкритого тексту

Роботу виконали:

Біла Анастасія і Лета Яна,
студенти 3 курсу ФТІ НТУУ «КПІ»,
спеціальність «Кібербезпека», група ФБ-02

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:

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку 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:

Обираємо текст «Проблеми розвитку психіки», автор Леонтьєв Олексій, текст написаний російською мовою.

Текст містить багато символів окрім власне літер, тому проводимо попередню фільтрацію тексту та створюємо два файли: один містить пробіли, інший - ні.

Створюємо функцію для підрахунку частоти букв у тексті, а також обчислення ентропії та надлишку. Результати частоти букв експортуємо у файл «frequency_of_letters.xlsx».

Для тексту з пробілами:

Частота букв (відсортована за спаданням частот):

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
частота	0,097366	0,084177	0,070685	0,060334	0,059946	0,053974	0,050193	0,042156	0,038166	0,029304	0,026881	0,025419	0,023973	0,022788	0,019951	0,018921	0,017591	0,01539

S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
з	ч	ь	г	б	й	х	ж	ю	щ	ш	ц	э	ф	ъ	ё
0,01539	0,01331	0,01326	0,01239	0,01153	0,01145	0,01025	0,00808	0,00649	0,00539	0,00458	0,00442	0,00414	0,00313	0,00066	2E-05

Ентропія:

H1: 4.026651479198996

Надлишковість:

R: 0.2017571617280558

Для тексту без пробілів:

Частота букв (відсортована за спаданням частот):

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
частота	0,11239	0,097166	0,081592	0,069643	0,069195	0,062303	0,057938	0,048661	0,044055	0,033826	0,031029	0,029342	0,027672	0,026304	0,023029	0,02184	0,020305	0,017764

S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
з	ч	ь	г	б	й	х	ж	ю	щ	ш	ц	э	ф	ъ	ё
0,01776	0,01537	0,01531	0,01431	0,01331	0,01322	0,01183	0,00933	0,00749	0,00622	0,00528	0,0051	0,00478	0,00361	0,00076	2,3E-05

Ентропія:

H1: 4.451702489104348

Надлишковість:

R: 0.11749510768391025

Створюємо функцію для підрахунку частоти біграм у тексті, а також обчислення ентропії та надлишку. Результати частоти біграм експортуємо у файл «frequency_of_bigrams.xlsx».

Для тексту з пробілами:

Частота біграм з перетином:

Ентропія:

Надлишковість:

Частота біграм без перетину:

а	б	в	г	д	е	ж	з	и	к	л	м	н	о	п	т	у	ф	х	ц	ш	щ	б	в	ю									
а	0.003422	0.001054	0.001583	1.91E-06	8.19E-05	0.001145	0	3.81E-06	0	0.000842	0.004861	0.000945	1.71E-05	1.91E-06	0.004838	0.000442	0	0.046666	7.62E-06	0.000297	0.006945	8.19E-05	0	0.0056	0								
а	0.001605	0.000692	0.001583	7.62E-06	5.91E-05	0	0	0.000621	0	0	1.91E-05	0.001707	0.000189	3.81E-06	0.000729	0.000447	0.000454	0.000191	0.000213	0	0.000747	1.91E-05	0	0	0.001655	2.1E-05							
а	0.002647	3.81E-06	0.000242	0.000147	0	45-005	0.000357	0	0	7.62E-06	0	0.000297	0	0.000118	0.000682	0	0.000447	0.000342	0	1.14E-05	0.000264	0	0.000587	0	0	0.002917	9.35E-06						
г	0.001108	0.000218	0.000189	0.000189	0.000695	1.91E-06	0.000203	0	0	3.81E-06	3.81E-06	5.72E-06	0	0.000206	0.001202	0.000189	0.000695	0.000206	0.000164	0.000145	0.000388	8.8E-06	0.000209	0.000186	0	0.000474	0						
д	0.001397	0.000416	3.81E-06	0.000189	0.000229	0	5.72E-06	0	0	1.91E-06	0.000288	0.000284	3.81E-06	0.000608	0.000101	3.81E-06	0.000219	0.000743	0.000174	0.000969	0.000606	0.000198	0.000773	0	0	0.000238	0						
д	0.006537	0.000356	7.62E-06	3.81E-06	0.000185	0.000692	0.000219	0.000151	0	0	2.1E-05	0.000348	0	1.91E-06	0	0.000215	0.000553	0.000143	0.000755	0.0002913	0.000161	0.000322	1.91E-06	0	4.57E-05	0.0001574	0						
д	0.006278	0.001286	0.000265	3.81E-06	0	0.000326	0.000565	0.000956	0	0	0	0.000341	0.000308	3.81E-06	2.1E-06	0	0.000379	0.000378	0.000283	0.000298	0.000192	0.000492	1.91E-06	0	0.000347	0.000358	3.81E-06						
д	0.001264	3.81E-06	1.91E-06	7.62E-06	1.52E-05	0.000452	3.81E-06	0	0	1.91E-06	0	0.000001	0.000238	8.19E-06	0.000648	0	0	0.0003178	0.0007475	0.0009998	2.1E-05	0.000295	0.001619	0.00049	9.53E-06	3.81E-06	0	0.0001509	1.91E-06				
ж	0.003742	0	0.001572	7.62E-05	1.14E-05	0.000271	0.000212	0	2.29E-05	0	0	0.000015	0	0	0.000286	3.81E-06	0	0	1.91E-05	0.000286	0.000207	2.29E-05	1.91E-06	0	0	0.000225	0						
ж	0.001591	3.81E-06	0	0.005238	0.0008484	5.91E-05	0.000676	1.91E-06	0	1.71E-05	0	9.3E-06	0	0	1.91E-06	1.52E-05	3.81E-06	0	0	0.000111	0.000205	0.000402	0.00206	0.000301	0.000211	0.000391	2.48E-05	0					
ж	0.003005	0.000148	0	0	0.000353	5.14E-05	0.000295	0	0	0.000939	0.000869	9.91E-06	0.000808	0	0	0	0	0	0.000494	0.000413	0.000192	0	0.000249	0.000482	0.000362	0.000447	0						
ж	0.012656	0.0383	1.91E-06	1.52E-05	0	0.0001616	0.0004459	5.14E-05	0	1.91E-06	0.003001	0.014909	6.0E-06	0.000207	0	0	0	3.81E-06	0.000187	0	3.81E-05	0.000296	2.9E-05	0.001262	4.76E-05	5.72E-05	1.71E-05	0.001511	1.91E-06				
ж	0.000233	5.72E-06	0	0	0	0.0001646	0	0	0	0	0	0.000265	0.000303	0.000466	4E-05	0	0	0	1.71E-05	0.0001179	4.57E-05	0.000501	0.000101	0.0001118	0.0001646	0.000993	5.14E-05	0.000107	0.000349	0			
ж	1.91E-06	0	0.000332	0.000362	0	1.91E-06	0	5.72E-06	0	0	0.000183	0.000406	1.91E-06	0.008787	1.91E-06	1.91E-06	0.004543	0	1.71E-05	0.0001115	5.72E-06	3.0E-05	0.000433	0	0	0	0.000663	0					
ж	0.006707	0.000277	0.000181	0.0004927	0	1.91E-06	0	7.62E-06	0	0	0.000897	0.0001771	0.000335	0.00128	0	0.0003742	0	0.000802	1.14E-05	0	0.000249	1.91E-05	0	1.91E-05	0.000355	5.72E-06	0.0003072	2.1E-05	0				
ж	0.001511	0.05093	0	0.002466	0.001446	3.81E-06	0.000739	0	0	0.002892	0.001227	0.000876	0.000985	7.62E-06	0.004887	0.000802	0.002452	1.14E-05	0.000444	0	0.000276	0.000145	1.91E-06	0.000183	0.000221	2.57E-05	0.00151	1.91E-06	0	0.0005498	0		
ж	0.004182	0.00363	3.83E-06	0	0	0.002092	0.00101	0	0	0	0.000194	0.000936	4.76E-05	1.14E-05	0	0	0.000109	0.000849	0.000303	8.38E-05	0	0	5.72E-06	0.000181	0.000109	1.14E-05	0	0	0.000161	0.000161	0.000109		
ж	0.009494	0.002037	0.00085	0.000217	0	0.000326	0.000037	0	0	0	0.0005084	0.0001033	0	0.000118	0	0.0001768	0.0001025	0.0001048	0.0001007	0.0001816	0	0	0.000276	0.000145	1.91E-06	0.000501	2.48E-05	0	0.0001604	0			
ж	0.012514	0.00688	0.0345E	0.002456	5.81E-05	0	0	0	0	0	0	0.0001991	0.000295	5.72E-05	1.91E-06	1.91E-06	1.24E-05	0.000837	0.00016181	0.000518	0	0.000234	1.14E-05	5.72E-06	0.0001114	0	0	0	0.0001605	0			
р	0.014643	0.000113	0.000568	0.000223	0.000497	0	0.000137	0.0003594	0	0	1.91E-06	2.48E-05	0	3.81E-06	1.91E-06	1.91E-06	2.48E-05	0	2.48E-05	0.0002633	5.14E-05	9.58E-06	0.000793	0	0.001075	0	1.91E-06	0	0.0000604	0			
ж	0.005661	0.000227	0.0002417	0.002357	5.72E-06	0	0.000892	0.0002908	0	0	0.0001833	2.1E-05	0.000476	5.72E-06	0	0.0002608	7.62E-06	1.91E-06	0.0005695	0.0001048	0.0004765	4.76E-05	0.0004516	0	0	0	0	0	0.0000402	0			
ж	0.014601	0.001343	0	0.0007056	0.0001616	0.0005915	0	7.61E-06	0	0	0.000654	0.0003536	7.62E-06	1.71E-06	1.91E-06	0.0005287	0	3.81E-06	6.0E-06	0.0008778	7.62E-06	0.0008439	0.002416	0.0003855	1.91E-06	0	0	0	0	0	0.0000402		
ж	0.006070	0.001772	0.001212	0.000236	0.0006787	1.91E-06	0.0004685	1.52E-05	0.000383	0	0.0008053	0.000233	0.000151	0	0	1.91E-06	0.000466	0.0005635	1.33E-05	0	0.000357	0.000207	0.000655	0.0005257	0.0001835	0	0	0	0	0	0.0001616	0	
ж	0.003323	0.003513	0.000229	0.000444	3.81E-06	0.002195	0.01334	1.91E-06	0.0001401	4E-05	0	0	7.62E-06	1.91E-06	0.000242	5.72E-06	0	0.000889	0.000483	0	0.00065	4.76E-05	0.005606	0.003955	5.91E-05	0	0	0	0	0	3.81E-06	0	
ж	0.001912	0.002048	3.81E-06	0.001625	0.000901	0.0001265	0.0000587	1.91E-06	3.81E-05	0	0.000114	0.001894	0	1.91E-05	0.0002028	7.62E-06	0	0.0002469	0.0000521	0.000335	2.04E-05	0	0.000275	1.91E-06	0	0	0	0	0	0.0002165	0		
ж	0.000648	9.72E-05	0.000634	0.000404	0	0.001255	0.000449	0	0.0001808	0	0.000396	0.001604	0.0001143	0	0.000277	0	0	1.91E-06	0	0.002841	2.86E-05	9.58E-06	0.005664	0.000345	0.0003072	0	0	0	0	0	0.000667	0	
ж	0.000442	0.000568	0	0.00101	0	0.001098	0.000558	0	0	0.001282	0.001218	0.000153	0.000192	0	0	0	0.0002464	3.81E-06	0.0002748	0	0.0007019	0.00132	0.000509	0.000509	0	0	0.000598	0.000137	0	0	0.0003083	0	
ж	0.000932	0.000651	0.000223	0.00195	0.000301	6.48E-05	0.0006589	1.91E-06	1.14E-05	7.62E-06	0.0002467	0.0004888	0.000355	4.95E-05	0	0.000246	0.0003241	7.62E-06	0.0001951	1.91E-06	0.0001743	0.000223	0.000889	0	0.000406	0.0001745	0	0	0	0	0	0.0001516	0
ж	0.000314	0.001947	1.91E-06	4.57E-05	0	0.001118	0.000688	7.62E-06	0	0	0.0000734	0.0002695	0.0001825	0	0	1.91E-06	0.0002462	0	0	0.000406	0.001916	0.000685	0.000888	0	5.91E-05	0	0	0.0001831	1.33E-05	0	0	0	0
ж	0.000324	1.91E-06	1.91E-06	0	0	8.76E-06	0.0001998	0	0.0003025	1.91E-06	0.000492	0.000459	7.62E-06	0.0001917	3.81E-06	0.000211	3.81E-06	0.00015	0	0.000621	1.91E-06	3.81E-06	5.72E-06	0	0.000339	9.58E-06	0	3.81E-06	0	0.0004021	0	0.0001301	0.0000402
ж	3.81E-06	0	0.000269	0.000189	1.4E-05	0	0	0.0001145	1.33E-06	0.0000579	0	0	0	1.91E-05	0.000248	0	0.000101	0.000183	0.000177	1.91E-06	0	0.000101	0.000183	0.000102	7.62E-06	0	4.38E-05	0	0	1.91E-06	0	0.0005667	0.0001404
ж	2.67E-05	0	0.000621	3.81E-06	0	0	0.000113	0.0000503	0	0.000101	0.000101	0.0003407	0.000388	0.00024	0	0.0000740	0.0003543	0.0005	0.001296	0	0.0001616	0.001202	7.62E-06	0.0000918	0	0	0.00162	0	0	0	0	0.0002328	

Ентропія:

Надлишковість:

H2: 3 9786843637874/01/

R: 0 217943307314739

Для тексту без пробілів:

Частота біграм з перетином:

Ентропія:

Надлишковість:

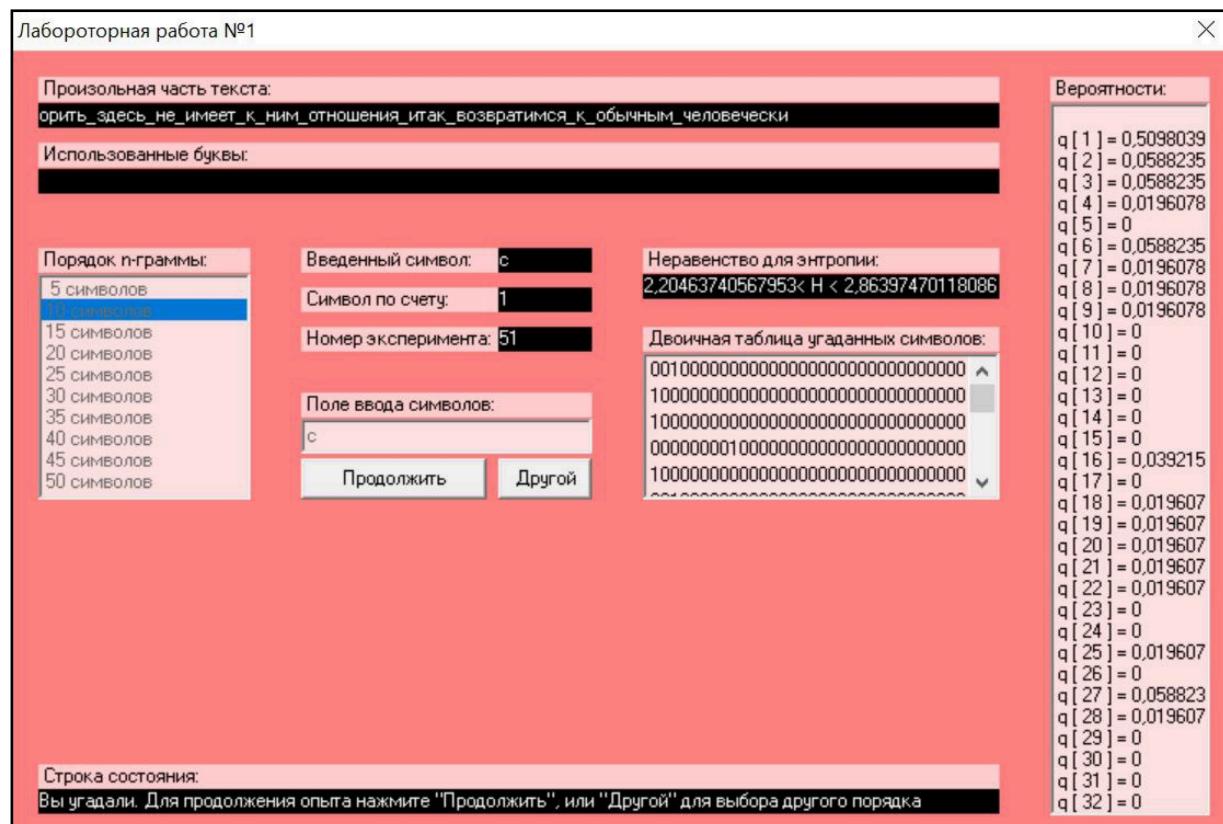
Частота біграм без перетину:

Ентропія:

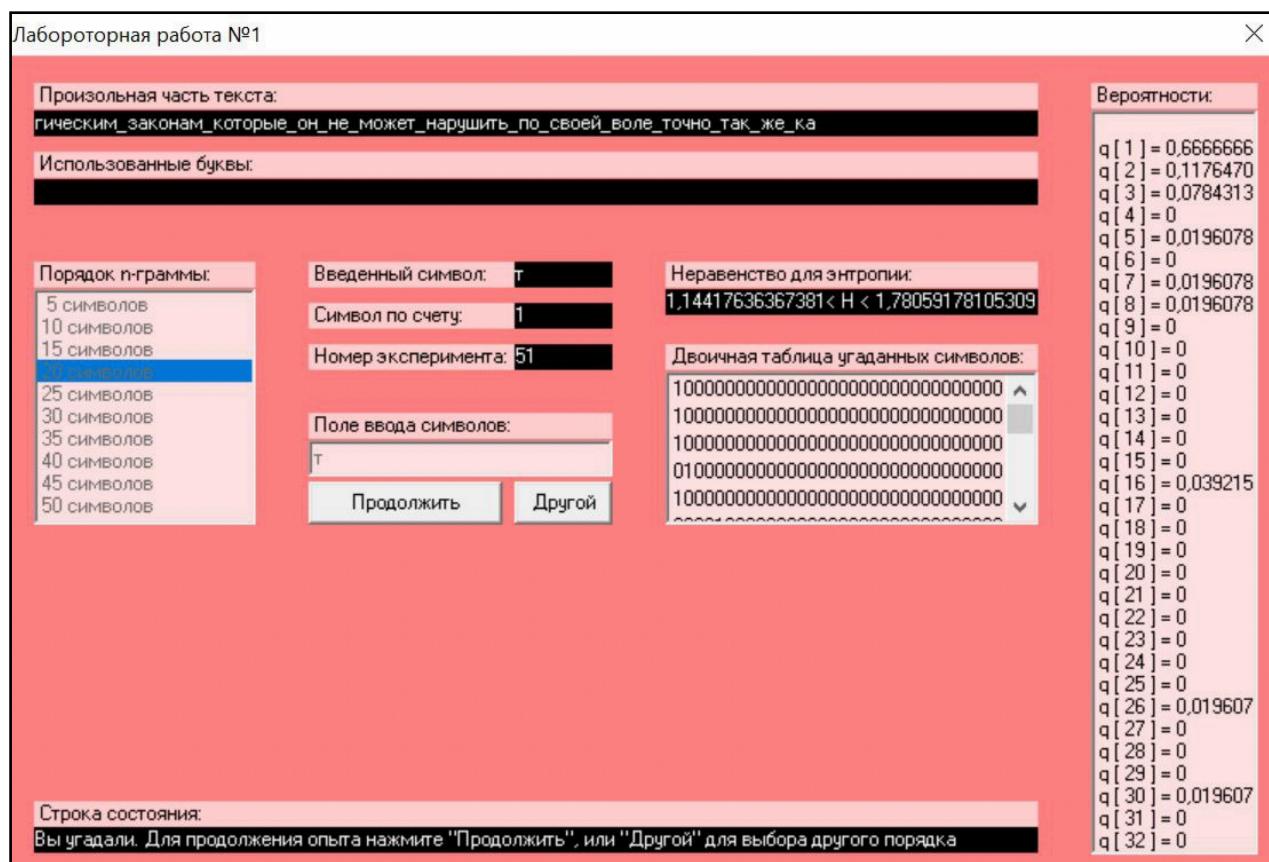
Надлишковість:

Використовуючи програму CoolPinkProgram, оцінимо значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$:

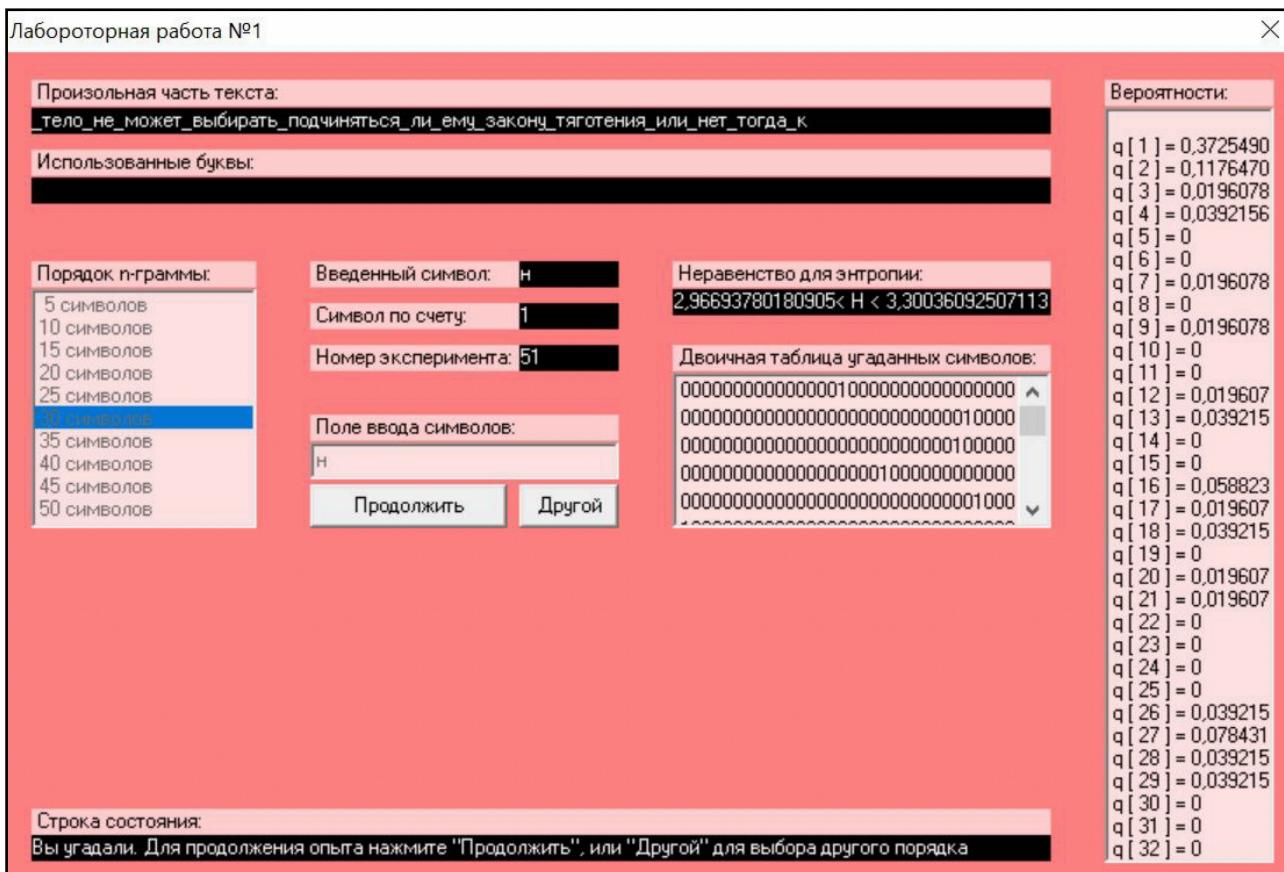
$$2.20463740567953 < H^{(10)} < 2.86397470118086$$



$$1.14417636367381 < H^{(20)} < 1.78059178105309$$



$$2.96693780180905 < H^{(30)} < 3.30036092507113$$



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

-для $H^{(10)}$: $0.432246047113959 < R < 0.562952982357390$;

-для $H^{(20)}$: $0.647015728961411 < R < 0.7731786342223144$;

-для $H^{(30)}$: $0.345736901800435 < R < 0.411834656133810$.

Висновки:

У результаті виконання комп’ютерного практикуму ми навчилися порівнювати різні моделі відкритого тексту, засвоїли поняття ентропії на символ джерела та його надлишковості за допомогою практичних навичок роботи з біграмами і літерами.