

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

КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела
відкритого тексту

Виконали:
студенти групи ФБ-32
Гереновська Мирослава
Клименко Іван

Мета роботи

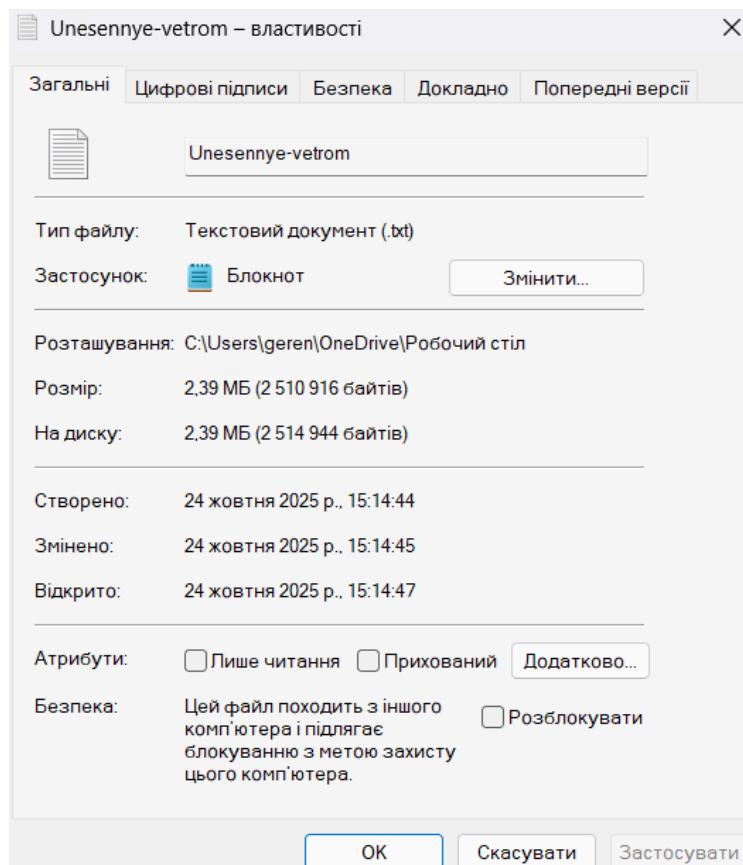
Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Порядок виконання роботи

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 біграмм, 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

Для виконання комп’ютерного практикуму ми обрали роман Маргарет Мітчелл “Віднесені вітром” (рос. “Унесённые ветром”) у форматі .txt як джерело відкритого тексту. Обсяг текстового файлу - 2,39 МБ.



Далі виконали фільтрацію тексту "Unesennye-vetrom.txt" відповідно до вимог практикуму: наша програма виконує переведення всіх прописних літер у стрічні, замінює букви "ё" на "е", "ъ" буквою "ъ" (що узгоджується з вимогами програми CoolPinkProgram), вилучення всіх символів, окрім літер російського алфавіту, з їх заміною на пробіли та нормалізацію послідовності пробілів до одного пробілу.

Зчитування файлу Unesennye-vetrom.txt
Довжина тексту (з пробілами): 2388762 (вимога: >1 Мб)

Після фільтрації тексту, ми сформували дві моделі джерела відкритого тексту для подальшого аналізу:

1. Текст "з пробілами": у цій моделі пробіл вважається окремим символом алфавіту. К-сть букв алфавіту $m=32$ (31 літера + пробіл).
Теоретична ентропія $H_0 = \log_2 32 = 5,0$ біт.
2. Текст "без пробілів": з тексту вилучено всі пробіли, він складається виключно з літер. К-сть букв алфавіту $m=31$.
Теоретична ентропія $H_0 = \log_2 31 \approx 4,954$ біт.

Збереження відфільтрованого тексту (з пробілами та без)

Завдання 1:

Частоти літер з пробілами (freq_h1_with_spaces.xlsx):

1	Символ	Кількість	Частота
2	<пробіл>	399960	0,167434
3	о	213529	0,089389
4	е	173460	0,072615
5	а	169403	0,070917
6	и	132933	0,055649
7	н	131227	0,054935
8	т	128206	0,05367
9	с	105335	0,044096
10	л	104496	0,043745
11	в	81990	0,034323
12	р	81443	0,034094
13	к	69172	0,028957
14	м	62349	0,026101
15	д	61926	0,025924
16	у	56516	0,023659
17	п	51912	0,021732
18	ь	40491	0,016951
19	я	39486	0,01653
20	ы	38123	0,015959
21	б	34872	0,014598
22	г	34612	0,014449
23	з	31776	0,013302
24	ч	29177	0,012214
25	ж	23368	0,009782
26	й	21587	0,009037
27	х	18609	0,00779
28	ш	17736	0,007425
29	ю	11395	0,00477
30	э	9205	0,003853
31	щ	6252	0,002617
32	ц	6080	0,002545
33	ф	2136	0,000894

Частоти літер без пробілів (freq_h1_no_spaces.xlsx):

1	Символ	Кількість	Частота
2	о	213529	0,107366
3	е	173460	0,087218
4	а	169403	0,085178
5	и	132933	0,066841
6	н	131227	0,065983
7	т	128206	0,064464
8	с	105335	0,052964
9	л	104496	0,052542
10	в	81990	0,041226
11	р	81443	0,040951
12	к	69172	0,034781
13	м	62349	0,03135
14	д	61926	0,031137
15	ү	56516	0,028417
16	п	51912	0,026102
17	ь	40491	0,020359
18	я	39486	0,019854
19	ы	38123	0,019169
20	б	34872	0,017534
21	г	34612	0,017403
22	з	31776	0,015977
23	ч	29177	0,014671
24	ж	23368	0,01175
25	й	21587	0,010854
26	х	18609	0,009357
27	ш	17736	0,008918
28	ю	11395	0,00573
29	э	9205	0,004628
30	щ	6252	0,003144
31	ц	6080	0,003057
32	ф	2136	0,001074

Частоти біграм з перетином та пробілами (freq_h2_overlapping_spaces):

1	Біграма	Кількість	Частота
2	а_	50098	0,020972
3	и_	49318	0,020646
4	о_	48502	0,020304
5	е_	47024	0,019686
6	_н	39914	0,016709
7	_с	39690	0,016615
8	_п	36685	0,015357
9	_в	36094	0,015111
10	то	28840	0,012073
11	_о	27526	0,011523
12	ь_	26345	0,011029
13	на	25907	0,010845
14	не	25094	0,010505
15	_и	24087	0,010083
16	я_	23178	0,009703
17	ст	23070	0,009658
18	ла	22730	0,009515
19	по	21668	0,009071
20	но	20599	0,008623
21	_к	20461	0,008566
22	ка	19845	0,008308
23	_д	18996	0,007952
24	_т	18738	0,007844
25	ли	18638	0,007802
26	ал	18527	0,007756
27	ни	18067	0,007563
28	ко	17878	0,007484
29	он	17747	0,007429
30	т_	17728	0,007421
31	_м	17542	0,007344
32	й_	17203	0,007202

Частоти біграм з перетином та без пробілів (freq_h2_overlapping_no_spaces.xlsx):

1	Біграма	Кількість	Частота
2	то	30062	0,015116
3	на	25996	0,013071
4	не	25256	0,012699
5	ст	23711	0,011922
6	ла	22838	0,011483
7	он	22538	0,011332
8	по	21669	0,010896
9	но	21011	0,010565
10	ен	20111	0,010112
11	ка	19964	0,010038
12	ал	19323	0,009716
13	ли	19132	0,00962
14	ни	18650	0,009378
15	ко	18619	0,009362
16	ов	18112	0,009107
17	ос	17994	0,009048
18	ет	17770	0,008935
19	от	16606	0,00835
20	ра	16313	0,008202
21	го	16092	0,008091
22	ть	15223	0,007654
23	ас	15150	0,007618
24	ер	15126	0,007606
25	во	14856	0,00747
26	ро	14748	0,007416
27	ло	14736	0,007409
28	та	14580	0,007331
29	ес	14338	0,007209
30	ан	13989	0,007034
31	ом	13827	0,006952
32	ол	13821	0,006949

Частоти біграм без перетину та з пробілами (freq_h2_non_overlapping_spaces):

1	Біграма	Кількість	Частота
2	а_	24898	0,020846
3	и_	24795	0,02076
4	о_	24361	0,020396
5	е_	23503	0,019678
6	_с	19896	0,016658
7	_н	19808	0,016584
8	_п	18460	0,015456
9	_в	18128	0,015178
10	то	14314	0,011984
11	_о	13800	0,011554
12	ь_	13211	0,011061
13	на	13051	0,010927
14	не	12627	0,010572
15	_и	12073	0,010108
16	ст	11593	0,009706
17	я_	11577	0,009693
18	ла	11341	0,009495
19	по	10779	0,009025
20	_к	10376	0,008687
21	но	10208	0,008547
22	ка	9846	0,008244
23	_т	9439	0,007903
24	_д	9350	0,007828
25	ли	9270	0,007761
26	ал	9233	0,00773
27	ни	8993	0,007529
28	ко	8874	0,00743
29	т_	8808	0,007375
30	он	8804	0,007371
31	_м	8692	0,007277
32	й_	8529	0,007141

Частоти біграм без перетину та без пробілів (freq_h2_non_overlapping_no_spaces):

1	Біграма	Кількість	Частота
2	то	15062	0,015147
3	на	13078	0,013152
4	не	12675	0,012746
5	ст	11838	0,011905
6	ла	11431	0,011495
7	он	11304	0,011368
8	по	10736	0,010796
9	но	10535	0,010594
10	ен	10013	0,010069
11	ка	10005	0,010061
12	ал	9644	0,009698
13	ли	9604	0,009658
14	ни	9397	0,00945
15	ко	9295	0,009347
16	ов	9121	0,009172
17	ет	8915	0,008965
18	ос	8863	0,008913
19	от	8266	0,008313
20	го	8090	0,008136
21	ра	8087	0,008133
22	ер	7656	0,007699
23	ас	7539	0,007581
24	ть	7526	0,007568
25	ло	7526	0,007568
26	ро	7406	0,007448
27	во	7404	0,007446
28	та	7299	0,00734
29	ол	7067	0,007107
30	ес	7047	0,007087
31	ан	6927	0,006966
32	ат	6859	0,006898

Матриця біграм з перетином та з пробілами (matrix_h2_overlapping_spaces.xlsx):

1-й символ	<пробирка>	а	б	в	г	д	е	ж	и	к	л	м	н	о	п	р	т	ф	ч	ш	щ	ы	ъ										
а	5141	16491	36004	8678	18996	64741	5476	1506	24087	21	20461	1688	17542	39914	27526	36865	9894	39690	18738	11988	1266	3988	677	12510	1754	221	0	0	7435	541	3355		
б	50098	18211	6370	1278	4833	1255	8839	6765	266	1080	10888	18527	7330	8710	3823	19014	9814	9815	11527	154	159	2765	394	1685	1603	543	0	0	2	1847	4839		
в	688	1865	9	148	12	24	5062	33	9	1817	0	350	1664	133	594	5785	0	2520	352	355	0	129	17	6	22	412	9025	328	14	37	1167		
г	13470	12733	12	83	37	711	9615	0	134	5770	0	488	1674	267	2050	14129	463	1240	7023	333	397	0	28	80	135	2102	8	6422	368	0	0	478	
д	1738	2248	0	19	33	3498	615	0	1	1666	0	367	3997	1	668	15975	0	209	28	58	1466	0	0	122	14	0	0	0	26	0	0		
е	2996	11730	90	1439	39	76	1087	1888	12	5010	0	566	1628	84	3920	9516	171	2199	583	219	4185	2	98	635	85	2	0	1446	1907	15	162	926	
ж	47024	112	2911	3233	7188	6019	5995	1882	3707	239	6549	2149	12411	8333	15955	444	2294	13847	9468	15724	353	24	1363	515	2326	1783	1486	0	0	367	450		
ж	743	3179	47	19	91	1854	9988	71	0	3417	0	255	63	17	2075	567	0	7	48	7	379	0	0	0	1	402	0	0	0	0	133		
з	3	1497	521	1759	878	1628	944	241	73	737	0	455	109	1046	1070	0	724	15	53	698	0	0	20	4	0	0	0	0	898	337			
и	49018	117	1116	4753	785	516	482	482	456	156	1763	533	1609	5664	7555	357	660	1735	9988	1916	46	4310	1738	1924	1598	360	0	0	0	525	2734		
й	17203	117	1116	4753	785	516	482	482	2	63	19	0	2	0	209	26	162	1011	33	0	12	704	644	0	15	0	47	630	146	1	0	0	58
к	10985	284	2	298	1	1	1222	37	33	675	0	45	1541	1	935	1978	0	4023	474	1108	3647	0	0	51	17	42	0	0	0	211	31		
л	8347	22730	61	175	267	701	13074	708	214	18638	0	1257	1540	86	845	14010	95	0	2012	360	297	7	3	0	314	2	2	1967	7763	52	2500	3800	
м	17187	6300	29	10	121	0	8906	0	0	7973	0	226	507	125	3156	8040	197	241	250	20	5507	8	4	86	5	2011	185	202	1	103			
н	7242	25097	8	17	344	857	25094	23	51	18067	0	1895	9	4	5947	20599	0	284	2160	1969	4605	273	5	825	363	12	548	7887	2579	0	306	3212	
о	48502	6	7914	13523	10978	12091	4331	4430	3707	2111	8746	4768	19785	11817	17747	42	2561	11032	13390	18853	118	155	1173	265	4267	2166	571	0	0	52	493	1867	
п	92	2726	0	0	0	0	0	0	0	2957	0	796	0	796	0	796	0	796	0	796	0	1	17	0	1	0	1	116	0	0			
р	18525	185	709	309	884	1030	12446	695	175	947	0	256	2511	553	1915	14601	267	314	369	1200	5720	279	339	319	351	281	434	1192	4604	255	2199		
с	8635	2776	257	3220	86	487	9883	135	9	3497	0	10367	1787	2018	2132	5577	3896	2070	2070	1396	24	409	92	111	1	547	8586	99	423	7476			
т	17728	14322	100	14509	21	208	11334	0	17	7864	0	1402	1642	88	2345	28840	248	5231	2556	5533	3243	8	23	170	392	25	62	3283	15223	3	247	1539	
у	14887	113	1885	2133	4076	419	4006	716	603	235	1890	3218	2719	628	66	1778	989	2658	3513	0	64	942	28	1822	2132	736	0	0	466	2165	191		
ф	44	127	0	0	0	0	463	0	0	289	0	0	97	4	6	278	0	539	69	16	116	11	0	0	0	0	23	3	48	0	3		
х	9559	30	810	517	9	2	65	0	0	434	0	2	514	107	429	4855	0	317	125	20	317	1	0	0	5	27	0	0	86	40	0		
ц	550	1238	0	181	0	0	2090	0	0	629	0	29	0	2	0	629	0	0	0	0	274	0	0	0	0	0	0	0	456	0	0		
չ	88	251	0	14	0	0	6868	0	0	3734	0	941	94	27	2040	99	0	34	0	8355	1587	0	0	0	0	0	0	0	682	0	0		
շ	277	2207	0	60	0	0	0	0	0	3597	0	1974	2551	46	404	424	0	1	210	624	0	0	1	0	0	0	0	122	2	0			
՛	55	857	0	0	0	0	0	0	0	187	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	0	0		
մ	12698	0	521	1975	300	313	3214	101	177	13	2669	423	4775	2655	408	0	347	558	1401	1640	8	0	2564	21	274	1056	21	0	0	0	0	18	
ն	26345	0	252	21	373	60	95	3	291	154	0	2553	0	537	2369	9	5	0	2666	255	0	4	0	272	175	1170	13	0	0	0	0	1224	1005
զ	4	0	0	18	31	9	0	0	176	0	385	52	663	239	523	0	13	180	35	5747	0	14	4	0	0	1112	0	0	0	0	0		
յ	5648	80	1233	1	159	895	4	182	53	1	16	52	236	73	136	0	4	187	311	906	0	8	35	4	150	168	807	0	0	0	55	2	
ր	23178	0	46	96	357	1525	232	535	579	40	125	453	2148	900	2121	0	183	257	1208	2917	2	411	0	477	169	224	27	411	0	0	0	196	214

Матриця біграм з перетином та з пробілами (matrix_h2_overlapping_no_spaces.xlsx)

1-й символ	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	x	y	ш	в	з	и	о			
																						x	y	ш	в	з	и	о			
а	579	3957	10817	2150	8874	4059	3260	7972	3342	1081	14376	19233	9579	13898	3696	6519	10064	15150	13741	1787	353	3104	481	3318	1790	580	0	0	967	1945	5556
б	1876	15	175	19	37	5090	34	31	1859	0	563	1669	154	623	5852	25	3529	385	48	2639	0	132	17	9	24	412	9025	338	385	37	1184
в	12791	401	631	581	1521	1066	180	1473	6374	0	1418	2013	666	3041	14856	1530	1686	8489	1411	1657	50	173	125	437	2195	26	6422	368	463	32	579
г	2778	130	168	54	3620	675	13	46	1809	0	439	3992	38	829	16092	155	2126	212	129	1534	3	13	4	164	20	0	0	51	0	18	
ж	11775	198	1681	135	218	10950	1311	93	5203	0	754	1678	167	4325	9716	416	2297	850	333	4283	11	122	637	128	98	0	1446	1907	74	164	1567
е	558	5147	7380	8213	8298	6960	2577	4762	2660	6550	4111	13328	10883	20111	3445	7395	15126	14338	17770	1818	128	1979	591	3707	1996	1518	0	0	716	417	878
ж	8188	26	68	96	1875	10005	63	56	3470	0	308	72	61	2178	616	52	23	93	44	393	0	7	143	1	0	0	133	26	0	28	
з	11543	404	2001	1032	1822	1027	269	280	863	0	704	720	951	4291	124	331	832	343	248	803	13	34	12	75	27	2	898	337	66	4	600
и	743	3218	9686	5526	5973	907	5815	4511	1762	6246	11626	7263	12101	4216	5390	287	12119	12205	1795	252	4731	1832	3038	1827	388	0	0	821	590	3466	
и	359	708	1406	551	1881	1000	377	402	1146	0	1215	349	904	2407	1002	1522	629	2438	1369	501	114	193	110	1108	29	21	0	0	220	36	186
и	1254	1123	1255	205	859	1459	500	357	1728	0	1611	1728	524	2081	1869	1681	217	1747	1748	1749	28	127	177	91	61	0	0	507	305	366	
и	22884	584	432	1078	1739	758	399	19132	1681	1676	562	14736	734	245	2799	654	3165	3165	611	61	11	734	37	4	1967	7765	179	255	3048		
и	5556	666	1529	573	990	9187	189	415	9159	0	1242	801	799	4804	9237	1769	571	2024	721	5978	56	171	55	870	70	9	2011	185	438	33	1270
и	25996	536	674	487	1158	2526	122	246	18650	0	2192	95	667	21011	800	423	1646	2251	6336	279	83	832	520	39	549	7887	2579	73	312	3273	
о	375	10237	18112	11890	12384	5757	5320	4393	4196	8766	13821	18287	22538	4789	2701	12208	17994	16606	1602	288	1764	329	6020	2345	593	0	0	943	948	2565	
и	2728	3	10	0	2	4242	1	1	2979	0	406	2221	1	351	2169	73	1367	20	1744	2	1	17	59	14	1	924	137	116	10	357	
и	16613	258	872	365	1104	12487	911	214	9611	0	928	4530	663	2088	14748	436	374	555	1266	5350	45	356	39	349	285	45	3134	1192	435	256	2234
и	2822	511	3800	272	853	9917	247	200	3765	0	10874	7323	2773	3155	5974	4562	770	2747	23711	1750	77	472	100	1017	140	2	647	8356	460	438	7600
и	14580	872	6243	538	996	12026	293	497	8896	1	215	1895	761	4137	9006	1860	5611	4222	6273	3816	41	154	197	1014	81	69	3283	15223	375	255	1870
и	426	1806	3134	2433	467	696	4292	978	1780	235	2735	3397	2395	9943	2903	1297	3599	4099	301	105	1064	44	2564	2178	740	0	0	655	2190	460	
и	129	2	1	0	2	465	0	0	290	0	1	98	5	8	79	4	539	76	18	116	11	0	0	2	0	0	23	3	61	0	3
и	1492	361	1238	270	524	249	148	210	1047	0	617	795	545	1187	5407	1044	656	1080	355	543	42	76	24	209	89	6	0	86	155	31	97
и	1246	9	243	8	25	245	5	8	807	0	457	5	22	57	68	461	47	10	3	1	17	0	3	1	456	0	6	1	9		
и	4252	4	18	1	14	6899	1	3	3744	0	949	75	288	2048	203	8	36	10	8357	1589	0	0	0	1	282	0	0	682	0	0	
и	2208	20	71	3	23	4206	3	3	3566	0	1886	2353	57	736	418	66	10	3	218	685	1	4	1	2	3	0	0	1130	6	1	
и	857	0	1	2	3054	0	0	1788	0	1	0	188	0	1	28	1	0	237	0	0	0	2	0	0	82	0	0	0	0		
и	178	1011	3153	462	903	3609	341	526	1024	2669	941	4927	3171	1771	989	1585	1776	2613	2218	387	35	2712	34	543	1111	25	0	0	200	18	
и	330	1145	2805	782	1303	2116	262	914	2008	0	3811	406	1623	4988	1864	2110	480	5275	1363	711	48	198	309	1260	1256	23	0	0	481	1246	1374
и	9	0	18	31	9	1	0	176	0	385	52	663	242	54	0	13	180	36	5747	0	14	4	0	0	1112	0	0	1	0		
и	158	1399	487	310	1170	121	314	175	430	33	418	368	311	590	320	480	333	855	1141	119	39	84	22	418	222	815	0	0	94	71	98
и	274	965	3883	794	2688	785	856	1289	1497	125	1745	2553	1772	4955	1332	2399	774	3493	3897	634	59	767	197	823	118	418	0	0	341	235	485

Зведені результати розрахунку ентропії та надлишковості тексту (lab1_results.xlsx):

Параметр	З пробілами	Без пробілів
H0 (теоретична)	5	4,954196
H1 (ентропія)	4,361644277	4,455914891
H2 (з перетином)	3,960822999	4,145655246
H2 (без перетину)	3,960501328	4,145543886
R (надлишковість за H1)	0,127671145	0,100577593
R (надлишковість за H2 з перетином)	0,2078354	0,163203223

Результати:

	Параметр	З пробілами	Без пробілів
0	H_0 (теоретична)	5.000000	4.954196
1	H_1 (ентропія)	4.361644	4.455915
2	H_2 (з перетином)	3.960823	4.145655
3	H_2 (без перетину)	3.960501	4.145544
4	R (надлишковість за H_1)	0.127671	0.100578
5	R (надлишковість за H_2 з перетином)	0.207835	0.163203

Висновки щодо наших результатів:

1. Експериментально підтверджено спiввiдношення $H_2 < H_1 < H_0$. Ентропiя бiграм ($H_2 = 3,960823$ бiт) менша за ентропiю окремих символiв ($H_1 = 4,361644$ бiт). Тобто це значить, що наявнi сильнi статистичнi зв'язки мiж сусiднimi лiтерами у мовi.
2. Значення ентропiї бiграм з перетином та без перетину є дуже близькими (3,960823 та 3,960501).
3. Розрахована нами надлишковiсть тексту на основi бiграм становить 0,207835 (це десь 20,8%) для тексту з пробiлами. Тобто це значить, що завдяки граматичнiй структурi, текст мiстить близько 20% зайвої iнформацiї.
4. Вилучення пробiлiв призвело до незначного пiдвищення ентропiї H_1 (з 4,361644 до 4,455915). Це значить, що пробiл є найчастiшим символом i його видалення робить розподiл iмовiрностей решти символiв бiльш рiвномiрним.

Завдання 2:

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

Результати для $H^{(10)}$:

$$2.8481 < H^{(10)} < 3.1692$$

Лабораторная работа №1

The screenshot shows the user interface of the CoolPinkProgram software. On the left, there is a sidebar with a list of options: 'Произвольная часть текста:' (Arbitrary part of the text), 'онимание_'; 'Использованные буквы:' (Used letters); 'Порядок n-грамм': dropdown menu with options 5 символов, 10 символов (selected), 15 символов, 20 символов, 25 символов, 30 символов, 35 символов, 40 символов, 45 символов, 50 символов; 'Символ по счету:' (Symbol by index); 'Номер эксперимента:' (Experiment number) set to 50; 'Строка состояния:' (Status line) empty. In the center, there are input fields: 'Введенный символ:' (Entered symbol), 'Символ по счету:' (Symbol by index), 'Номер эксперимента:' (Experiment number) set to 50, 'Поле ввода символов:' (Input field for symbols) containing 'Продолжить' (Continue) and 'Другой' (Other) buttons. To the right, there is a large text area for results: 'Неравенство для энтропии:' (Inequality for entropy) showing '2.84818123429229 < H < 3.16920882454116'; 'Двоичная таблица угаданных символов:' (Binary table of guessed symbols) showing a 4x8 grid of binary digits; 'Вероятности:' (Probabilities) showing a list of probabilities for each symbol from 1 to 32, with q[1] = 0.3673469 and q[32] = 0.102040.

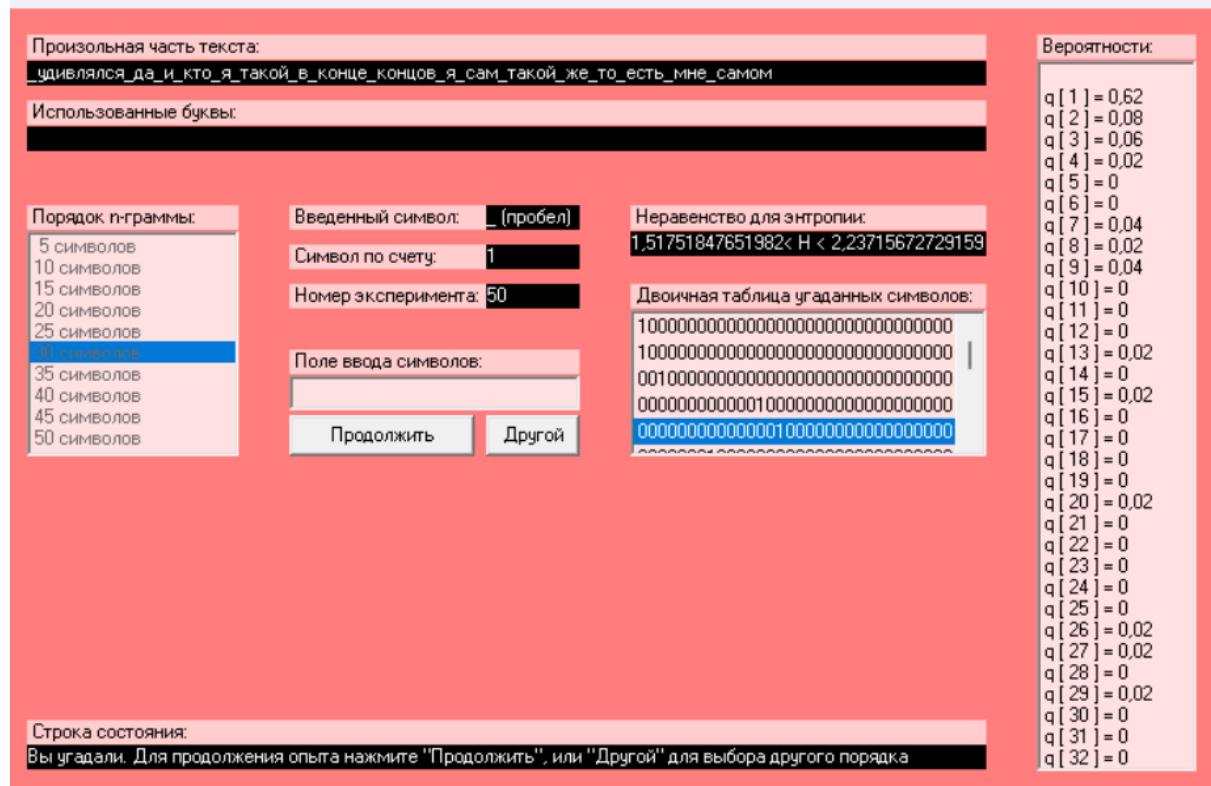
Результаты для $H^{(20)}$:

$$2.3434 < H^{(20)} < 3.9879$$

Результати для $H^{(30)}$:

$$1.5185 < H^{(30)} < 2.2371$$

Лабораторна робота №1



Короткий висновок щодо наших результатів:

Ми помітили зменшення ентропії: найнижчі значення ентропії отримано при $H^{(30)}$ ($1.5185 - 2.2371$). Це каже про те, що зі збільшенням к-сті відомого тексту невизначеність наступного символу суттєво зменшується.

Завдання 3:

Використовуючи отримані значення ентропії, оцінили надлишковість російської мови в різних моделях джерела:

Для цього використали формулу: $R = 1 - \frac{H_{\infty}}{H_0}$, $H_0 = 5.0$ біт ($H_0 = \log_2 32 = 5.0$ біт) $H^{(10)}$:

$$\text{Нижня межа: } R = 1 - \frac{2.8481}{5} \approx 1 - 0.56962 \approx 0.43038 \approx 43,04\%$$

$$\text{Верхня межа: } R = 1 - \frac{3.1692}{5} \approx 1 - 0.63384 \approx 0.36616 \approx 36,62\%$$

$H^{(20)}$:

Нижня межа: $R = 1 - \frac{2.3434}{5} \approx 1 - 0.46868 \approx 0.53132 \approx 53.13\%$

Верхня межа: $R = 1 - \frac{3.9879}{5} \approx 1 - 0.79758 \approx 0.20242 \approx 20.24\%$

$H^{(30)}$:

Нижня межа: $R = 1 - \frac{1.5185}{5} \approx 1 - 0.3037 \approx 0.6963 \approx 69.63\%$

Верхня межа: $R = 1 - \frac{2.2371}{5} \approx 1 - 0.44742 \approx 0.55258 \approx 55.26\%$

Висновки до наших розрахунків:

- Ми помітили, що чим більший обсяг тексту (контекст) враховується, тим вища його надлишковість.
- Розрахунок за допомогою python-коду показав близько 20% надлишковості, оскільки він базується суто на частоті пар літер. Натомість метод CoolPinkProgram виявив надлишковість до 70%. Думаємо, що це через те, що людина використовує розуміння змісту та граматики тексту, які недоступні для простого частотного аналізу.
- Отримане нами значення надлишковості приблизно 69.63% майже збігається з еталонними даними для російської мови (76%). Це свідчить про те, що текст "Унесенные ветром" є типовим прикладом природної мови.

Table 3. Entropy and redundancy of fiction texts for several languages by Piotrovski *et al.*

	Language	\hat{H} bits	\hat{H} bits	\hat{R} (%)	\hat{R} (%)
1	Russian	1.19	0.70	76	86
2	Polish	1.29	0.83	74	84
3	English	1.10	0.65	77	86
4	German	1.36	0.83	71	82
5	French	1.36	0.78	71	84
6	Rumanian	1.26	0.78	74	84
7	Kazakh	1.35	0.81	75	85

Покликання на знайдену нами інформацію:

<https://gwern.net/doc/cs/algoritm/information/compression/1994-levitin.pdf>

Висновки:

У ході виконання комп'ютерного практикуму 1, нами було проведено експериментальну оцінку ентропії та надлишковості джерела відкритого тексту (на прикладі російської мови, а саме на романі Margaret Mitchell "Віднесені вітром" (рос. "Унесённые ветром")) за допомогою різних моделей.

Результати:

	Параметр	З пробілами	Без пробілів
0	H_0 (теоретична)	5.000000	4.954196
1	H_1 (ентропія)	4.361644	4.455915
2	H_2 (з перетином)	3.960823	4.145655
3	H_2 (без перетину)	3.960501	4.145544
4	R (надлишковість за H_1)	0.127671	0.100578
5	R (надлишковість за H_2 з перетином)	0.207835	0.163203

Експериментально підтвердили, що врахування залежностей між символами суттєво зменшує невизначеність (ентропію) тексту. Встановили суттєву різницю між автоматичними методами (python-код) та методом передбачення (CoolPinkProgram): Python враховує лише "механічні" зв'язки між сусідніми літерами, тому дає оцінку надлишковості лише 20%, а CoolPinkProgram дозволяє оцінити реальну ентропію мови, оскільки людина використовує знання граматики та логічного змісту (цей метод показав надлишковість на рівні 69.63%). Отримане нами значення надлишковості приблизно майже збігається з еталонними даними для російської мови. Це свідчить про те, що текст "Унесенные ветром" є типовим прикладом природної мови.