

НТУУ "КПІ ім Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела
відкритого тексту

Виконали:

студенти групи ФБ-14

Разумний Ілля

Болгов Микола

Перевірила:

Селюх П.В.

Київ 2023

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 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

Обраний текст міститься у файлі [ds.txt](#)

Програма [razik_bolgov.py](#) на мові Python3

Частота букв в тексті з пробілом:

Letter frequencies in the text with spaces		
" "		0.15158885306697925
"о"		0.09107583374326358
"а"		0.07048301060088764
"е"		0.06929070790229701
"и"		0.06027894762828683
"н"		0.051603665416455606
"т"		0.050129745209302155
"с"		0.04580124752898923
"л"		0.04331606445017839
"в"		0.04018535552381771
"р"		0.03990556670037543
"к"		0.030471380163521022
"м"		0.02629832071845381
"п"		0.026206886462426922
"д"		0.025910639472899802
"у"		0.024559241168822382
"ы"		0.016580687987916048
"я"		0.01600282348982611
"ь"		0.015013504839615171
"б"		0.014501473005864593
"г"		0.01425642919971253
"з"		0.013969325635788099
"ч"		0.012707532902617032
"й"		0.010573457366949443
"ж"		0.008642365879661548
"х"		0.007815800205178471
"ш"		0.00720684806003939
"ю"		0.0045808562269471385
"ц"		0.004368728752964755
"щ"		0.0028765216946059274
"э"		0.0020883584076541446
"ф"		0.001709820587702824

Ентропія з пробілом:

$H_1 = 4.403827131527339$

Надлишковість з пробілом:

$R.1 = 0.11923457369453216$

Частота букв в тексті без пробіла (ми не враховуємо пробіл як частину алфавіта):

Letter frequencies in the text without spaces

"о"		0.10734870006423161
"а"		0.08307647872812784
"е"		0.08167114276230424
"и"		0.07104921693472947
"н"		0.060823888987080396
"т"		0.059086617839145074
"с"		0.05398473098162286
"л"		0.051055510770650035
"в"		0.04736542614873283
"р"		0.04703564638988158
"к"		0.03591581778913926
"м"		0.03099714190875662
"п"		0.03088937074573334
"д"		0.030540192177537902
"у"		0.02894733438805378
"ы"		0.019543222702642117
"я"		0.01886210895233497
"ь"		0.017696024968423048
"б"		0.017092506455492665
"г"		0.016803679738590267
"з"		0.01646527828669716
"ч"		0.014978036236975856
"й"		0.012462657292012432
"ж"		0.010186530328960698
"х"		0.00921227901523022
"ш"		0.008494523069495157
"ю"		0.0053993352674664725
"ц"		0.005149306169252456
"щ"		0.0033904807887124794
"э"		0.002461493363451781
"ф"		0.00201532074853539

Ентропія без пробіла:

$H.1 = 4.467203167107593$

Надлишковість без пробіла:

$R.1 = 0.0982991211426687$

Частота біграм, що перетинаються, в тексті з пробілами:

bigram_table_cross_spaces.xlsx

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я	
а	7.31E-06	0.000739	0.003059	0.00066	0.002183	0.000997	0.001287	0.002864	0.000139	0.000892	0.003502	0.008564	0.003365	0.004113	1.1E-05	0.001918	0.003526	0.00376	0.005528	0.000165	0.000263	0.001317	0.00032	0.000913	0.000966	0.000227	0	0	1.28E-05	0.000719	0.001986	0.016487
б	0.000975	1.1E-05	4.39E-05	3.66E-06	1.1E-05	0.002461	3.11E-05	7.31E-06	0.000936	0	9.33E-05	0.001037	3.84E-05	0.000258	0.002092	1.28E-05	0.001317	0.00015	1.28E-05	0.001009	0	3.11E-05	1.65E-05	7.13E-06	3.47E-05	0.000163	0.002791	0.000432	0	7.86E-05	0.000223	0.000225
в	0.006258	1.28E-05	2.93E-05	2.38E-05	0.000415	0.000565	1.83E-06	0.000336	0.003703	0	0.000287	0.000808	9.87E-05	0.000956	0.000411	0.000161	0.000816	0.001966	0.000358	0.000739	0	0.000106	6.95E-05	7.13E-05	1.28E-05	0.000172	8.96E-05	0	0	0.000274	0.006786	
г	0.001489	1.83E-06	1.46E-05	0	0.000724	0.000316	0	1.83E-06	0.000695	0	0.000139	0.000135	7.31E-06	0.000219	0.000847	3.66E-06	0.001487	0	3.47E-05	7.31E-06	0.000658	0	0	9.14E-06	1.1E-05	0	0	0	3.66E-06	0	0.000552	
д	0.000485	3.84E-05	0.001315	1.1E-05	4.39E-05	0.000544	9.69E-05	1.46E-05	0.003361	0	0.000318	0.000525	0.000123	0.001642	0.000886	0.000152	0.000927	0.000441	8.59E-05	0.001434	0	4.57E-05	0.000424	3.11E-05	7.13E-05	0	0.000578	0.000426	0	3.47E-05	0.000408	0.00115
е	0.000169	0.001019	0.002734	0.002379	0.002964	0.001754	0.000983	0.001335	0.002144	0.002219	0.001372	0.006183	0.002306	0.006969	0.002019	0.00079	0.006239	0.004771	0.004736	0.00207	4.39E-05	0.000594	0.000556	0.000977	0.000664	0.000649	0	0	0.000106	0.000197	0.015234	
ж	0.001432	3.29E-05	0	0.146E-05	0.000772	0.003165	7.31E-06	0	0.001483	0	0.000194	2.38E-05	7.31E-06	0.000931	6.4E-05	0	3.66E-06	1.28E-05	3.66E-06	0.00024	0	0	1.83E-06	4.75E-05	0	0	0	0	2.56E-05	0	0	0.000181
з	0.005298	0.000106	0.000817	0.000183	0.000794	0.000561	6.77E-05	1.1E-05	0.000499	0	9.51E-05	0.000216	0.000318	0.001421	0.000667	3.66E-06	0.000192	3.84E-05	1.28E-05	0.000346	0	7.31E-06	6.58E-05	7.31E-06	0	0.000353	0.000124	0	1.83E-06	0.000271	0.001492	
и	0.00023	0.000342	0.00229	0.000614	0.001573	0.001922	0.002074	0.002213	0.000722	0.001452	0.002862	0.004945	0.002355	0.003981	0.000552	0.001165	0.000159	0.003259	0.004758	5.49E-05	9.33E-05	0.001478	0.000991	0.002631	0.000466	0.00034	0	0	1.83E-06	0.000291	0.00103	0.016334
й	0	5.49E-06	0	1.83E-06	0.000159	5.49E-06	0	1.65E-05	0	0	0.00017	0.000124	5.49E-05	0.000375	1.28E-05	0	1.83E-06	0.000521	0.000421	0	1.83E-06	1.83E-06	5.12E-05	0.000199	9.14E-05	5.49E-06	0	0	0	0	0	0.008353
к	0.006506	0	0.000366	3.66E-06	9.14E-06	0.000726	1.1E-05	8.23E-05	0.003826	0	2.74E-05	0.000856	2.01E-05	0.000395	0.008615	7.31E-06	0.00207	0.000243	0.000693	0.001818	0	1.83E-06	0.00019	1.83E-06	1.1E-05	0	0	0	3.66E-06	0	0.003988	
л	0.004883	6.58E-05	2.74E-05	0.000155	4.94E-05	0.004661	0.00024	2.38E-05	0.007783	0	0.000492	0.000287	7.31E-06	0.00039	0.005806	6.95E-05	1.83E-06	0.001686	0.000143	0.001403	7.31E-06	0	0	0.000148	3.66E-06	3.66E-06	0.000653	0.00425	0	0.000832	0.001609	0.007637
м	0.003535	0.000294	0.000171	1.65E-05	0	0.003367	0	1.83E-06	0.003195	0	0.000113	0.000174	9.87E-05	0.001103	0.003354	0.000177	0.000101	0.000148	1.83E-05	0.001841	4.02E-05	1.46E-05	3.11E-05	4.21E-05	1.1E-05	3.66E-06	0.000852	0.00011	0	0	0.000402	0.007137
н	0.008984	1.83E-05	1.65E-05	0.000159	0.000708	0.007328	0.00143	3.29E-05	0.006786	0	0.000479	3.66E-06	0	0.002679	0.009445	1.83E-06	6.77E-05	0.000742	0.000922	0.00273	2.01E-05	5.49E-06	0.000596	0.000267	9.14E-06	5.85E-05	0.004135	0.001081	9.14E-06	0.000104	0.001086	0.002986
о	1.83E-05	0.003348	0.008	0.00399	0.005201	0.001543	0.001876	0.001196	0.000722	0.003767	0.00203	0.006673	0.005499	0.00508	0.00179	0.00141	0.006821	0.0067	0.005385	9.33E-05	0.000133	0.000625	0.000154	0.001831	0.000898	0.000267	0	7.86E-05	0.000216	0.000569	0.016771	
п	0.002017	0	0	0	1.83E-06	0.002425	0	0.00141	0	6.4E-05	0.000836	5.49E-06	0.000137	0.009655	0.001004	0.005824	7.13E-05	0.000124	0.000849	5.49E-06	0	7.31E-06	1.46E-05	1.1E-05	5.49E-06	0.000158	0.000508	0	0.000174	0.000364	0.000936	
р	0.007838	9.87E-05	0.000369	0.00036	0.000315	0.005696	0.000347	0.000115	0.005062	0	0.000413	0.000108	0.00036	0.00039	0.00134	0.007847	0.000115	5.67E-05	0.000582	0.000706	0.002675	5.12E-05	0.000179	7.86E-05	8.78E-05	0.000179	5.49E-05	0.000158	0.000508	0.000174	0.000894	0.001776
с	0.001812	6.58E-05	0.001304	4.02E-05	0.000176	0.002933	9.14E-06	7.31E-06	0.002017	0	0.004345	0.000206	0.000795	0.000958	0.002644	0.001772	0.000254	0.00096	0.012014	0.000653	1.65E-05	0.000174	6.03E-05	0.000294	7.31E-05	0	0.000419	0.002913	0	0.000252	0.00367	0.002851
т	0.0067	1.65E-05	0.002776	5.49E-06	0.000214	0.005883	0	1.83E-06	0.00382	0	0.00056	0.000216	2.01E-05	0.001282	0.009646	7.68E-05	0.00357	0.000962	6.4E-05	0.002244	5.49E-06	2.38E-05	0.000113	9.69E-05	9.14E-06	3.66E-05	0.001565	0.004467	0	0.000126	0.000576	0.002565
у	8.59E-05	0.001064	0.000503	0.000744	0.001659	0.000249	0.001264	0.0004	1.83E-05	0.000139	0.001022	0.002264	0.000942	0.000399	2.38E-05	0.000728	0.000992	0.001126	0.001346	5.49E-06	3.84E-05	0.000528	1.83E-05	0.000812	0.00068	0.000225	0	0	1.46E-05	0.001024	8.23E-05	0.0061
ф	0.000236	0	0	0	0	0.000457	0	1.83E-06	0.000289	0	1.83E-06	7.31E-05	7.31E-06	1.1E-05	0.000309	0	8.78E-05	2.01E-05	2.19E-05	8.41E-05	1.65E-05	0	0	0	0	0	1.1E-05	2.56E-05	0	3.66E-06	0	5.85E-05
х	0.000735	0	0.000181	3.66E-06	0	0.000113	0	0	0.000294	0	9.14E-06	0.000183	0.000135	0.00023	0.000193	0	0.000132	6.4E-05	3.29E-05	0.000172	0	5.49E-06	1.83E-06	0	1.28E-05	0	0	1.83E-06	7.31E-06	0	0.003308	
ц	0.00949	0	0.00015	0	0	0.001024	0	0	0.000763	0	0.000205	0	5.49E-06	0	0.000192	0	0	5.49E-06	0	0.000274	1.83E-06	1.83E-06	1.83E-06	0	0	0	0.000276	0	0	0	0.000519	0
ч	0.002209	0	7.31E-06	0	0	0.003308	1.83E-06	0	0.001671	0	0.000675	6.22E-05	3.66E-06	0.000858	0.001011	0	5.3E-05	0	0.00175	0.000574	0	0	0	0	7.13E-05	0	0	0.000216	0	0	0.001147	
ш	0.001022	0	3.11E-05	0	0	0.001993	0	0	0.001622	0.000461	0.000518	1.28E-05	0.000386	0.00305	2.56E-05	0	1.83E-06	1.83E-06	0.000139	0.000331	0	1.1E-05	0	0	0	0	0	0.000247	5.49E-06	0	9.33E-05	0
щ	0.000464	0	1.83E-06	0	0	0.001225	0	0	0.000828	0	0	0	0	5.67E-05	0	0	0	0	0	0	0.000143	0	0	0	0	0	0	0	0	3.29E-05	0	0.000124
ы	0	0.000161	0.000865	0.000144	0.000141	0.001379	3.29E-05	6.22E-05	4.21E-05	0.00198	0.000223	0.001858	0.001333	0.000207	0	0.000205	0.000628	0.000675	0.000618	1.83E-06	0	0.000132	5.49E-06	0.000132	0.000448	1.83E-05	0	0	0	1.46E-05	0.004623	0
ь	0	8.41E-05	2.01E-05	0.00019	0.000157	0.000658	0	0.000104	1.68E-05	0	0.000834	0	0.000196	0.0013	6.03E-05	5.49E-06	3.66E-06	0.000888	0.00019	0	3.29E-05	5.3E-05	9.69E-05	5.3E-05	0.000545	1.83E-05	0	0	0.000395	0.000913	0.008339	0
э	0	0	0	0	0	0	1.83E-06	0	0.000443	0	1.83E-06	8.78E-05	0.000144	8.05E-05	2.56E-05	1.83E-06	3.47E-05	6.58E-05	4.75E-05	0.001551	0	2.01E-05	2.01E-05	0	0	1.83E-06	0	0	0	0	0	0
ю	3.66E-06	0.000304	5.49E-06	1.65E-05	0.000315	0	2.93E-05	4.75E-05	0	5.49E-06	0.000165	2.38E-05	8.59E-05	6.58E-05	0	7.31E-06	7.13E-05	0.000104	0.000399	0	0	1.28E-05	3.11E-05	0.000213	2.56E-05	0.000413	0	0	0.329E-05	1.83E-06	0.002293	0
я	0	0.000108	0.00247	9.51E-05	0.00041	9.51E-05	0.000172	0.000199	3.29E-05	0.000108	0.000201	0.000885	0.0004	0.000955	0	0.000106	0.000197	0.000594	0.001417	1.83E-06	5.49E-06	0.000177	7.13E-05	0.000203	2.19E-05	0.000256	0	0	0.000113	8.78E-05	0.000851	0
	0.002348	0.006554	0.014895	0.00444	0.006914	0.002335	0.001867	0.004901	0.009346	1.46E-05	0.0																					

Ентропія для біграм, що перетинаються, з пробілами:

H.2 = 4.021115482275827

Надлишковість для біграм, що перетинаються, з пробілами:

R.2 = 0.19577690354483457

Частота біграм, що не

Частота біграм, що перетинаються, в тексті без пробілів:

bigram_table_cross_not_spaces.xls

	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ	ы	ь	э	ю	я	
а	0.000371	0.001636	0.005557	0.001323	0.003406	0.001502	0.001711	0.003975	0.001414	0.001052	0.005521	0.010374	0.005009	0.006473	0.001136	0.004528	0.004636	0.006568	0.007408	0.000679	0.000481	0.001681	0.000418	0.001461	0.001257	0.000287	0	0	0.000328	0.000856	0.002541	
б	0.001155	1.29E-05	6.9E-05	8.62E-06	1.94E-05	0.00291	3.66E-05	1.72E-05	0.001142	0	0.000119	0.001226	4.96E-05	0.003325	0.002498	2.8E-05	0.001356	0.000919	2.37E-05	0.001201	4.31E-06	3.66E-05	1.94E-05	4.31E-05	0.000192	0.003289	0.005059	3.02E-05	9.27E-05	0.000269		
в	0.007458	0.000393	0.000463	0.000399	0.000875	0.000746	8.62E-05	0.000608	0.000466	2.16E-06	0.001121	0.001101	0.000405	0.001711	0.000822	0.001048	0.001043	0.003519	0.000976	0.00106	3.88E-05	0.000175	0.000144	0.000239	0.001013	2.37E-05	0.003686	0.000106	0.000175	4.31E-06	0.000562	
г	0.001772	2.8E-05	7.11E-05	1.29E-05	0.000892	0.000382	6.47E-06	0.002877	0	0.000491	0.000655	0.001777	0.002084	0.004664	0.000293	0.001138	0.000655	0.000181	0.001735	2.16E-06	6.9E-05	0.0005	1.37E-05	9.7E-05	0	0	0.000679	0.000502	1.51E-05	4.1E-05	0.000463	
д	0.005076	0.000103	0.001688	4.53E-05	0.000108	0.005953	0.000121	6.47E-05	0.002865	0	0.000491	0.000655	0.001777	0.002084	0.004664	0.000293	0.001138	0.000655	0.000181	0.001735	2.16E-06	6.9E-05	0.0005	1.37E-05	9.7E-05	0	0	0.000679	0.000502	1.51E-05	4.1E-05	0.000463
е	0.000394	0.002201	0.004865	0.003317	0.004397	0.002356	0.001272	0.002218	0.001088	0.002608	0.002552	0.007068	0.004559	0.009689	0.001319	0.00294	0.007077	0.007462	0.006415	0.000819	0.000116	0.000877	0.000691	0.001608	0.000953	0.000782	0	0	0.000241	0.000125	0.000286	
ж	0.001688	4.53E-05	2.16E-05	2.16E-05	0.000916	0.003785	8.62E-06	1.29E-05	0.001755	0	0.000241	3.02E-05	1.51E-05	0.001127	8.62E-05	2.8E-05	1.72E-05	3.23E-05	1.29E-05	0.000285	6.47E-06	2.16E-06	2.16E-06	8.62E-05	0	0	0	3.02E-05	2.16E-06	0	6.47E-06	
з	0.006266	0.000196	0.001121	0.000304	0.001073	0.000677	9.91E-05	0.000112	0.000629	0	0.000259	0.000272	0.000466	0.00177	0.000892	0.000244	0.000261	0.0002	9.91E-05	0.000442	1.08E-05	2.16E-06	2.16E-05	0.000123	1.72E-05	0	0	0.000416	0.000147	1.51E-05	2.16E-06	0.00033
и	0.000502	0.001276	0.004684	0.001216	0.002701	0.002578	0.000453	0.003289	0.002037	0.001714	0.004414	0.006158	0.003268	0.006466	0.002011	0.003604	0.001936	0.005891	0.006389	0.000558	0.000213	0.001884	0.001248	0.003524	0.000685	0.000429	2.16E-06	0	0.000198	0.000351	0.00136	
й	0.00019	0.00041	0.000696	0.000422	0.000767	0.000101	0.000188	0.000336	0.000632	0	0.000981	0.000336	0.000498	0.001039	0.000528	0.001048	0.000304	0.001638	0.000907	0.000203	7.76E-05	0.000101	0.000119	0.000053	0.000209	1.29E-05	0	0	0.905E-05	1.51E-05	8.41E-05	
к	0.007753	0.000237	0.000929	0.000134	0.000246	0.000914	0.000112	0.000237	0.004813	0	0.000269	0.001067	0.000183	0.00086	0.010473	0.000444	0.002539	0.000817	0.001071	0.002231	6.47E-06	3.66E-05	0.000241	0.000127	3.02E-05	0	0	0	8.19E-05	6.47E-06	5.6E-05	
л	0.005912	0.000375	0.001088	0.000446	0.000358	0.005708	0.000319	0.000261	0.009949	2.16E-06	0.001073	0.000429	0.000209	0.001198	0.000819	0.000877	0.0002	0.002755	0.000466	0.001806	1.72E-05	5.6E-05	1.72E-05	0.000461	6.04E-05	1.08E-05	0.000769	0.005009	8.62E-05	0.000981	0.001964	
м	0.004337	0.000677	0.000942	0.000364	0.000399	0.004076	0.000123	0.000259	0.004358	0	0.000629	0.000325	0.000511	0.00197	0.004404	0.001147	0.000345	0.000987	0.000321	0.002349	8.62E-05	9.27E-05	5.39E-05	0.000248	0.000103	6.47E-06	0.001004	0.000129	9.27E-05	4.31E-06	0.000567	
н	0.010622	0.000198	0.000403	0.000256	0.000994	0.008695	0.000207	0.000164	0.008195	0	0.000737	3.88E-05	0.000129	0.003464	0.011353	0.000466	0.000172	0.001287	0.002109	0.003352	2.59E-05	4.31E-05	0.000707	0.000401	4.31E-05	7.11E-05	0.004873	0.001274	3.02E-05	0.000123	0.001291	
о	0.000259	0.004975	0.011508	0.005229	0.007042	0.002162	0.002548	0.00214	0.001776	0.004444	0.003406	0.008139	0.007197	0.007719	0.001561	0.003636	0.008598	0.010629	0.000627	0.000231	0.000907	0.000278	0.002634	0.000198	0.000325	0	0	0.000358	0.000261	0.000866		
п	0.002393	7.98E-05	3.91E-06	2.59E-05	3.66E-05	0.002869	1.08E-05	3.66E-05	0.001705	0	0.000129	0.000996	3.02E-05	0.000263	0.01146	0.001332	0.006904	0.000207	0.00017	0.001035	6.47E-06	8.62E-06	8.62E-06	3.66E-05	1.94E-05	6.47E-06	0.000358	7.54E-05	2.16E-05	4.31E-06	0.000524	
р	0.009266	0.000209	0.000649	0.000485	0.000453	0.006731	0.000422	0.000198	0.006143	2.16E-06	0.000586	0.000168	0.000584	0.001759	0.009389	0.00362	0.001373	0.001345	0.014357	0.000916	4.53E-05	0.000237	7.76E-05	0.00039	9.7E-05	0	0.000494	0.003434	7.11E-05	0.000302	0.004347	
с	0.001219	0.000222	0.001832	0.00016	0.000358	0.003505	0.000121	0.00011	0.002509	0	0.005371	0.002789	0.001104	0.001459	0.003304	0.002466	0.000373	0.001345	0.014357	0.000916	4.53E-05	0.000237	7.76E-05	0.00039	9.7E-05	0	0.000494	0.003434	7.11E-05	0.000302	0.004347	
т	0.007999	0.000328	0.000382	0.000136	0.002542	0.007022	9.91E-05	0.000157	0.004789	0	0.009188	0.000317	0.001101	0.001938	0.011672	0.000688	0.0041	0.001714	0.000319	0.002746	2.59E-05	8.62E-05	0.00014	0.000272	5.6E-05	4.31E-05	0.001845	0.005266	9.91E-05	0.000151	0.000718	
у	0.000246	0.0015	0.001332	0.001054	0.002242	0.000384	0.001584	0.000634	0.000668	0.000166	0.001716	0.002724	0.001481	0.001203	0.00045	0.001625	0.001231	0.002054	0.001933	0.000172	7.54E-05	0.000685	3.66E-05	0.001188	0.000083	0.000272	0	0	9.05E-05	0.001211	0.00016	
ф	0.002191	4.31E-06	6.47E-06	2.16E-06	6.47E-06	0.000541	0	2.16E-06	0.000345	0	4.31E-06	8.84E-05	1.08E-05	1.51E-05	0.000364	4.31E-06	0.000106	2.59E-05	2.8E-05	0.000101	2.16E-05	0	0	0	0	0	0	1.29E-05	3.02E-05	0	4.31E-06	0
х	0.000963	0.000168	0.000535	0.000125	0.000175	0.000175	6.47E-05	0.0001	0.000612	0	0.000282	0.00033	0.000347	0.000281	0.00047	0.0003	0.0005	0.000181	0.00031	2.37E-05	2.37E-05	1.94E-05	6.04E-05	5.17E-05	2.16E-06	0	0	2.16E-06	5.6E-05	0	2.8E-05	0
ц	0.001123	6.47E-06	0.000235	1.72E-05	1.94E-05	0.001211	0	1.29E-05	0.000931	0	0.000265	6.47E-06	2.37E-05	2.8E-05	0.000246	3.66E-05	1.72E-05	4.74E-05	1.72E-05	0.000325	0.000233	6.47E-06	2.16E-06	4.31E-06	2.16E-06	2.16E-06	0.000325	0	4.31E-06	0	0	0
ч	0.002625	6.04E-05	0.000153	1.51E-05	5.17E-05	0.000301	4.31E-06	7.54E-05	0.002033	0	0.000851	9.91E-05	3.45E-05	0.001136	0.000231	0.000194	9.91E-05	0.000107	0.000293	0.000726	2.16E-06	2.16E-06	0	4.53E-05	8.84E-05	2.16E-06	0	0.000254	1.51E-05	0	1.51E-05	0
ш	0.001207	4.31E-06	4.96E-05	6.47E-06	2.16E-06	0.002349	0	0	0.00192	0	0.000552	0.00061	1.72E-05	0.000463	0.000366	3.66E-05	4.31E-06	1.51E-05	0.000168	0.000394	0	1.72E-05	4.31E-06	4.31E-06	2.16E-06	0	0	0.000291	0	6.47E-06	2.16E-06	0
щ	0.000554	4.53E-05	1.94E-05	1.08E-05	2.16E-06	0.001444	2.16E-06	4.31E-06	0.000985	0	6.47E-06	0	2.16E-06	6.9E-05	6.47E-06	4.31E-06	2.16E-06	8.62E-06	6.47E-06	0.000172	2.16E-06	0	0	2.16E-06	0	2.16E-06	0	2.16E-06	0	0	0	0
ы	6.47E-05	0.00042	0.001597	0.000323	0.000431	0.001742	7.98E-05	0.000274	0.000558	0.002334	0.000513	0.002357	0.001791	0.000787	0.000291	0.000877	0.00044	0.001331	0.000935	0.00136	2.37E-05	0.000142	2.37E-05	0.000237	0.000055	2.59E-05	0	0	5.6E-05	2.16E-06	5.17E-05	0
ь	0.000185	0.000044	0.001073	0.000444	0.000616	0.009974	7.98E-05	0.000455	0.000752	0	0.001584	0.000123	0.000662	0.000464	0.000653	0.000996	0.000308	0.001722	0.000679	0.000216	6.9E-05	0.000153	0.000134	0.000994	0.000694	3.23E-05	0	0	0.000151	0.00047	0.001173	0
э	0	0	0	4.31E-06	0	0	2.16E-06	0	2.16E-06	0	2.16E-06	0.000103	0.000165	8.62E-05	3.02E-05	2.16E-05	4.1E-05	7.76E-05	5.6E-05	0.001828	0	2.37E-05	2.37E-05	0	0	0	0	0	0	0	0	0
ю	4.53E-05	0.000448	0.000276	9.48E-05	0.000494	3.66E-05	0.000101	0.000127	0.000185	6.47E-06	0.000399	9.91E-05	0.000228	0.00025	0.000136	0.000315	0.000188	0.000351	0.000591	5.6E-05	3.45E-05	3.88E-05	4.31E-05	0.000207	6.25E-05	0.000489	0	0	2.59E-05	4.1E-05	4.1E-05	0
я	0.00017	0.000494	0.001543	0.000405	0.000935	0.000321	0.000317	0.000556	0.00072	0.000129	0.000938	0.001226	0.000793	0.000229	0.000623	0.001242	0.00047	0.00164	0.000297	0.000267	4.96E-05	0.000321	0.000106	0.000483	9.27E-05	0.000304	0	0	0.00011	0.00014	0.000	

Надлишковість для біграм, що не перетинаються, без пробілів:

$$R.2 = 0.15972178275825843$$

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

Лабораторная работа №1



Произвольная часть текста:
дочное_по

Использованные буквы:
р,

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: р
Символ по счету: 1
Номер эксперимента: 51
Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
 $1.88684736429788 < H < 2.67663484475755$
Двоичная таблица угаданных символов:
00000100000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
00000000000000000000000000000000
00000001000000000000000000000000

Вероятности:
q[1] = 0.5
q[2] = 0.12
q[3] = 0.08
q[4] = 0.02
q[5] = 0.06
q[6] = 0.02
q[7] = 0.02
q[8] = 0.04
q[9] = 0.02
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0.02
q[17] = 0
q[18] = 0.02
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0.02
q[26] = 0.02
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0.04

Строка состояния:
Вы не угадали. Введите другую букву

Лабораторная работа №1



Произвольная часть текста:
еальностью_если_бы_нацисты_не_знали_в_глубине_своего_сердца_так_же_хорошо_к

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: н
Символ по счету: 1
Номер эксперимента: 50
Поле ввода символов:
н

Продолжить Другой

Неравенство для энтропии:
 $1.36678530179095 < H < 2.17972183240961$
Двоичная таблица угаданных символов:
00000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
00100000000000000000000000000000

Вероятности:
q[1] = 0.6
q[2] = 0.16
q[3] = 0.02
q[4] = 0.02
q[5] = 0.02
q[6] = 0.04
q[7] = 0.02
q[8] = 0
q[9] = 0.02
q[10] = 0
q[11] = 0.02
q[12] = 0.02
q[13] = 0.02
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0.02
q[19] = 0
q[20] = 0
q[21] = 0.02
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

Произвольная часть текста:
е_какие_важные_уроки_слушая_что_ссорающиеся_говорят_друг_другу_они_говорят_н

Использованные буквы:

Порядок n-граммы:

- 5 символов
- 10 символов
- 15 символов
- 20 символов
- 25 символов
- 30 символов**
- 35 символов
- 40 символов
- 45 символов
- 50 символов

Введенный символ: **т**

Символ по счету: **1**

Номер эксперимента: **50**

Неравенство для энтропии:
 $1,46373049643962 < H < 2,10006166784613$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
00010000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Поле ввода символов:
т

Продолжить Другой

Вероятности:

q[1] = 0,62
q[2] = 0,1
q[3] = 0,02
q[4] = 0,04
q[5] = 0
q[6] = 0,02
q[7] = 0
q[8] = 0
q[9] = 0,04
q[10] = 0,08
q[11] = 0,02
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0,02
q[17] = 0
q[18] = 0,02
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0,02

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

Маємо наступні значення:

$$1.88684736429788 < H^{(10)} < 2.67663484475755$$

$$1.36678530179095 < H^{(20)} < 2.17972183240961$$

$$1.46373049643962 < H^{(30)} < 2.10006166784613$$

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

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

$H_0 = \log_2(32) = 5$ (32 букви алфавіту, враховуючи пробіл)

Для $H^{(10)}$:

$$R = 1 - \frac{1.88684736429788}{5} = 0.622630527140424$$

$$R = 1 - \frac{2.67663484475755}{5} = 0.46467303104849$$

Для $H^{(20)}$:

$$R = 1 - \frac{1.36678530179095}{5} = 0.72664293964181$$

$$R = 1 - \frac{2.17972183240961}{5} = 0.564055633518078$$

Для $H^{(30)}$:

$$R = 1 - \frac{1.46373049643962}{5} = 0.707253900712076$$

$$R = 1 - \frac{2.10006166784613}{5} = 0.579987666430774$$

Для підсумку маємо:

$$0.46467303104849 < R^{(10)} < 0.622630527140424$$

$$0.564055633518078 < R^{(20)} < 0.72664293964181$$

$$0.579987666430774 < R^{(30)} < 0.707253900712076$$

Труднощі при виконанні роботи:

Найскладнішим моментом при виконанні лабораторної роботи було побороти прокрастинацію та всістись за виконання завдань

На початку, основним завданням було відфільтрувати текст, ми використали Regular Expressions для швидкої обробки. Для аналізу тексту беруться тексти із пробілами та без них, проте ця проблема була вирішена доволі швидко шляхом модифікації функцій і додавання логічного параметру `spaces`. Так само були модифіковані функції `letter_frequency()` та `bigram_frequency()`, в останній також додано параметр `cross`, що відповідає за те, чи є біграма тою, у якій пари букв перетинаються, чи ні. Шляхом модифікації було скорочено код

Також, проблемою було правильно вивести та зберегти таблицю частот біграм. Вирішено проблему було за рахунок створення датафрейму pandas та його імпорту у Excel таблицю

Варто зазначити, що ми створили 2 функції для літер та біграм, які використовували для різних типів біграм та пробілів тексту, що робить код більше компактним. Окрім частот, вони також виводять підраховні ентропії та надлишковості для цих типів

Висновки:

Під час виконання лабораторної роботи ми ознайомились із наступними термінами:

Ентропія - міра невизначеності, показує кількість необхідної інформації для однозначного опису ансамблю (пара з множини символів та множини ймовірностей). Вирахувавши частоту появи символів алфавіту можна визначити ентропію тексту, тобто скільки інформації містить текст

Сукупна ентропія говорить нам, скільки інформації містять два ансамблі (із урахуванням взаємних залежностей)

Умовна ентропія говорить нам, скільки інформації залишиться в ансамблі X, якщо поведінку ансамблю Y буде однозначно визначено

На початку роботи для аналізу та підрахунку частоти та ентропії букв і біграм було обрано текст “Дванадцять стульев”. Було написано програму на мові Python3 *razik_bolgov.py*, що підраховує частоту букв і частоту біграм в тексті, а також ентропії H_1 та H_2 з надлишковістю

Було встановлено, що у російській мові (без пробілу) найчастіше трапляються літери “о”, “а”, “е”, “и”, “н”, “т”, “с”, “л”, “в”, “р”. Також було встановлено частоту різних біграм. Ці знання допомогли нам при експерименті з підрахунку надлишковості російської мови - при роботі із CoolPinkProgram, якщо ми не могли визначити наступну літеру із попередніх то вводили літери методом брутфорсу (від самої популярної), в більшості таких випадків вгадування літери було доволі швидким

В ході виконання лабораторної роботи при роботі із CoolPinkProgram ми також стикнулись із поняттям **умовної ентропії джерела**, що визначає, скільки інформації про наступний символ ми матимемо із значень попередніх. Було оцінено умовну ентропію при $n=10, 20, 30$. Завдяки чому було підраховано **надлишковість** російської мови (величину можливого ущільнення тексту деякою схемою кодування символів без втрати його змісту)

