

КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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.

Виконали Шанідзе Давид та Тивонюк Володимир

Текст, який був використаний для роботи: “Над пропастью во ржи” Джерома Девіда Сэлинджера

Хід 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 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 $(10) H$, $(20) H$, $(30) H$.
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Опис труднощів, що виникали, та шляхів їх розв'язання

Перш за все треба було знайти текст, який би ми аналізували. Знайти було не так важко, як скопіювати його весь у текстовий файл. Після цього цікавим моментом було редагування файлу з текстом, тобто видалення усіх зайвих символів та подвійних пробілів. Та складніше за все було впоратися з другим, але нам все-таки вдалося знайти досить хитрий вихід з даної ситуації (можете побачити у файлі edit.py). Далі прийшла черга виконання самої суті завдання. Якщо перше завдання було виконано досить просто та друге також не викликало запитань, то з третім вже довелося розбиратися. Після декількох годин гуглення та перечитування методички ми зрозуміли, що відповідь лежала майже перед нашими очима. Тоді ми закінчили з цим завданням й перейшли до третього. Там вже треба було працювати з даним нам екзе файлом. Найскладніше тут було вгадувати початок слова (тобто, що йшло після пробілу) та й взагалі робота тут була досить нудна й однообразна. У четвертому і вже останньому завданні проблемою було визначити H_∞ та H_0 . Але потім ми розібралися з даною проблемою, результати чого можна бачити далі.

Таблиці частот букв і біграм тексту, одержані значення Н1 та Н2

Таблиця Частоти Біграм:

(intersection, spaces)

	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	ь	э	я						
а	2.9e-05	0.0006337	0.0009213	0.0008615	0.0052102	2.9e-05	0.0012427	0.0004464	0.0001805	8.3e-06	0.0075212	0.0049782	0.0031559	0.018704	5.8e-05	0.0011597	0.0073762	0.0015076	0.000581	2.9e-05	0.0001077	0.0004224	0.0004837	0.0015345	0.0009795	0.0002589	0	4.1e-06	0						
б	0.0004033	7.87e-05	1.66e-05	0	5.38e-05	0.0012628	8.3e-06	0.0001804	0.0009153	5.59e-05	0	1.45e-05	6.82e-05	0	0.00039582	6.42e-05	0.0073765	0.0001199	0	0	0.0001657	1.66e-05	0	0	0.0001657	1.66e-05	0	0.0003313	1.24e-05						
в	0.0001394	4.1e-06	2.9e-05	0	0.0006736	0.0013067	0	0.0006275	0.0010903	0	0.0001015	2.9e-05	0	1.24e-05	0.0006608	0	0.000263	0.00010458	0.001783	0.0002982	0	0.0001015	1.24e-05	0	0.31e-05	0	0.0009712	2.40e-05	2.40e-05	0.0001242					
г	0.0003976	0	0	4.1e-06	0.0003013	0	0.0002562	0.0006233	0	0.47e-05	0	0.0002089	0.0004966	0	5.8e-05	1.66e-05	4.1e-06	0.0010665	0	0	0	0	0	0	0	0	0.94e-05	0.000116	0.000125	2.48e-05					
д	0.0020813	2.07e-05	0.0006061	0.0002239	0.0001069	0.0004307	0.0001599	0.0002236	0	0.0002516	0	0.0003334	0.0001597	0.0002155	0.000294	3.22e-05	0.0021199	0	0	0	0	0	0	0	0	0.00e-05	0.0001056	0.0003636	0.0002131						
е	0.0026525	0.0016877	0.0007068	0.0002625	0.0009347	0.0014931	0.0032015	0.0002559	0.0012632	1.24e-05	0.0005623	0.0032118	0.0034769	0.0113399	0.0001596	0.0015925	0.0004561	0.0004345	0.0009542	0.0001154	0.0001898	4.14e-05	0.0006015	0.0041271	0.0028045	0.0013005	0.0006896	0.0005384	0	0.9e-05					
ж	0.0012011	8.3e-06	0	0.94e-05	0.0008345	1.24e-05	0.0001505	0.0001454	0	0.0001595	0	4.1e-06	0.0017498	0	0.0001945	0	0.0001719	1.24e-05	0	0.0013522	0	0	0	0	0	0	2.9e-05	0	1.24e-05	3.31e-05					
з	0.0035781	2.07e-05	0.0002609	0	0.0009949	0	0.0001636	0.0001651	0	3.73e-06	0.0001167	1.24e-05	0	0	0.0001945	0	0.0001167	1.24e-05	0	0.0001583	0	0	0	0	0	0	3.31e-05	0.000116	1.24e-05	8.3e-06	0.0001201				
и	4.97e-05	0.0009581	0.002574	0.0004784	0.0023152	0.0001339	0.0001186	0.0002174	0	0.0004984	0.0070532	0.0010037	0	0.0001803	0.0001867	0.0001083	0.0001589	0.0038014	1.24e-05	0.0002444	0.0002154	0.0001512	0.0017229	0.0012549	0.0004349	4.1e-06	2.07e-05	0	4.1e-06	2.48e-05					
й	0.000791	0	0	0	0.0022522	0	0.0001958	0	0.0001958	0	0	0.001435	0	0	0.001435	0	0	0.001532	0.0004597	0.00029675	0.0000352	0.0000352	0	0	0	0	0.0014848	0	2.07e-05	2.48e-05	8.3e-06				
к	0.0006157	1.00e-05	0.0001284	4.76e-05	0.0002161	0.0007228	0.0002589	0.97e-05	0.0001959	5.93e-05	3.52e-05	0.0003844	0.0001263	0.000631	0.00061	0.0001532	0.0004597	0.00029675	0.0000352	0.0000352	0.0000352	0	0	0.0001387	0.0002754	0.0001996	0	0.0001429	0.0013605	3.31e-05	8.3e-06	0.0003479			
л	0.0003356	0.0004360	0.000251	0.0006564	0.0006751	0.0057858	2.48e-05	0.0001698	0.0012561	7.45e-05	0.0009446	0.0007579	4.97e-05	0	0.0001983	0.0006936	7.45e-05	0.0027083	0.0001305	0.0002336	1.66e-05	5.8e-05	0	0	0	0	0.41e-06	0.0000943	0	0.0001242	0.0001429	0.0013605	3.31e-05	8.3e-06	0.0003479
м	0.0013129	1.66e-05	7.45e-05	0	5.18e-05	0.0033195	4.1e-06	8.28e-05	0.0015513	7.87e-05	0	2.1e-06	1.66e-05	0	0.0002454	0	0.0002816	0.0008718	2.48e-05	0.0014019	4.1e-06	8.3e-06	0	0	0	0	2.1e-06	8.3e-06	0	0.0004084	0.0002216	8.3e-06	4.97e-05	0.0003148	
н	0.0001385	0.0007973	0.0001387	0.0001583	0.0058332	0.0000949	0.0017164	0.0020781	0.0006772	0.0004204	0.0001595	0.0002408	0.001137	0.0073307	4.76e-05	0.0011245	0.0011224	0.0009239	0.0001822	0	0.0001480	0	0.0003119	0.0004618	1.24e-05	0.0001615	0.0000532	0.0001657	4.14e-05	0.0004349	0	0.0001242			
о	4.14e-05	0.0015215	0.0004961	0.0074777	0.0032116	0.0010877	0.0001261	0.0002361	0	0.0002371	0.0057596	0.0001539	0.0004445	0.0004017	0.0001818	0.002311	0.0019221	0.0145392	1.84e-05	0	0.0001843	0.0026941	0.0001594	0.0002582	0.0001586	4.1e-06	0	8.3e-06	0	0.0001125	0	0.0001242			
п	0.0008049	0	7.66e-05	1.24e-05	2.28e-05	0.0005591	0	6.2e-06	0.0004899	1.66e-05	0	1.24e-05	8.49e-05	8.3e-06	0.0009029	2.48e-05	7.04e-05	0.0017126	8.7e-05	0.0004017	0	0	0	0	0	0	0	0.00147	8.3e-06	0	0.41e-05	0			
р	0.0008037	0.0006856	0.0005715	0.0007393	0.0012157	0.0004406	0	0.0001636	0.0004784	0	0.001551	0	3.73e-05	0.0003645	0.0003395	0.0052578	4.76e-05	0.0003438	0.0027521	0.0007682	8.28e-05	6.83e-05	0	0	0	0	0.000234	0	9.73e-05	2.28e-05	4.1e-06				
с	0.0043213	2.9e-05	0.003466	1.24e-05	0.0002195	0.0032814	1.34e-06	2.9e-05	0.0015238	0.0003914	0.0001347	0.0014371	8.28e-05	0.0006501	0.0004938	0.0001835	0.0001429	0.0006005	0.0010409	0.0009919	0	0	0	0	0	0	0	0.0001681	0.0009029	8.3e-06	0.0003562	0.0001225			
т	0.0059321	0	0.0003127	6.63e-05	0.0002324	0.0006316	8.3e-06	2.48e-05	0.0004628	0.0007289	0.0007952	8.49e-05	5.8e-05	0.0003439	0.0001482	5.8e-05	0.0021238	0.0008051	0.0002692	0.0019631	8.7e-05	4.76e-05	0	0.0001655	0.0001201	0.0001480	0.0004784	0.0001284	0.0024391	0.000676	0.0018078	0.0001242			
у	9.93e-05	0.0012756	0.0004452	0.0005322	0.0001934	5.18e-05	0.0005115	0.0004929	0	0.0002147	0.0017813	0.0010870	0	0.0017126	0.0001281	0.0005467	0.0023869	0.0003499	0.0014235	8.3e-06	6.63e-05	0.000147	0.0002257	0.0005811	0.0012549	3.73e-05	0	0	0	0	0				
ф	7.97e-05	0	0	4.1e-06	4.1e-06	0.0001491	0	0.0001134	0	0	6.63e-05	0	6.63e-05	0	0.0001835	0	0.0001935	0.00007673	0.0014522	0	1.24e-05	0	0	0	0	0	0	0	7.04e-05	0	0.0001242	0			
х	0.0008059	2.28e-05	1.24e-05	0	3.52e-05	0.0004121	0	0.0014082	0	4.1e-06	0	0	2.48e-05	0.0005343	0	0.0001835	0.0001835	8.3e-06	0.0002071	0	0	0	0	0	0	0	0	0.0005943	0	2.48e-05	7.04e-05	0			
ц	0.0001894	4.1e-06	1.24e-05	0	0.0001274	0.0003498	4.1e-06	0	0.0006192	1.41e-05	4.1e-06	0	0.0004307	6.21e-05	0	0.41e-05	2.28e-05	5.8e-05	2.48e-05	0	0	0	0	0	0	0	0	4.1e-06	3.52e-05	0	2.48e-05	0			
ч	0.0007579	8.3e-06	0.0002485	4.1e-06	1.24e-05	0.001896	0.0001367	0	0.0001284	2.07e-05	0.0001693	0.0021681	4.1e-06	0.0001693	0.0021681	4.1e-06	0.0001693	0.0021681	0.0001255	0.0001323	1.24e-05	0	0	0	0	0	0	4.1e-06	0	9.11e-05	0.0001429	0			
ш	0.0012073	0	0	0	6.63e-05	0.0013083	0	4.1e-06	0.0008635	2.07e-05	0	2.9e-05	4.1e-06	8.3e-06	0.0012715	8.3e-06	0.0002326	7.45e-05	0	0.0009339	1.66e-05	4.1e-06	0	0.0001325	0	0	0	0.0003439	0.0004349	2.07e-05	4.1e-06	0			
щ	0.0001367	0.0004991	1.24e-05	0	0	0.00088785	0	0	0.94e-05	0	0	0	0	0.58e-05	0.0001815	0	1.04e-05	0	0	6.42e-05	0	0	0	0	0	0	0	0	5.18e-05	4.1e-06	0	8.7e-05	0.0002073		
ь	0	0.000484	0.0019582	0	0	0.0001367	0	0.0009112	0.0000381	0.0001267	0.0000912	0.0000381	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912	0.0001267	0.0000912			
э	0	0.0001815	0.0001222	0	0.0008014	0	1.66e-05	0.0001325	0	0	0.003274	8.49e-05	0.0014827	0	0.0001515	0.0003186	0.0024663	0.0027264	0	0	0	0	0	0	0.0003186	0.0001335	2.48e-05	0	0	0	0	0	0		
я	0.0001714	8.3e-06	0.3e-06	0	0	8.3e-06	0	0	1.66e-05	0	0	1.24e-05	0	0	0.0003583	7.87e-05	4.1e-06	0.3e-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
а	0.0014961	0.0006939	0.0001698	0	0	0.0001698	0	0	0.0001698	0	0	0.0001698	0.0001183	0.002135	0.0006709	0.0004307	0.001023	0.000342	0.000352	1.3e-06	0	0	0	0	0	0	0	1.24e-05	0.0002536	0	0.184e-05	0			

(intersection, NO spaces)

	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	ь	э	я		
а	4.51e-05	0.0009585	0.0076552	0.0013397	0.0001828	4.51e-05	0.0022157	0.0002929	0.0001578	1.29e-05	0.0116699	0.0077241	0.0004981	0.0164648	9.02e-05	0.0018035	0.0114715	0.0023445	0.0102348	4.51e-05	0.0001675	0.0006657	0.00076	0.0021864	0.0005233	0.0004026	0	6.4e-06	0.64e-06	я	
б	0.0006026	0.0001224	2.50e-05	0	8.73e-05	0.0002189	1.29e-05	0.0002931	0.0014235	8.7e-05	0	0.0001758	0	0.0004757	0	9.90e-05	0.0001224	9.66e-05	0.0008083	0	0	0	0	0	0	0.0002576	2.50e-05	0	0.0000513	1.93e-05	0
в	0.0004813	6.4e-06	0.51e-05	0	0.0001274	0.0004987	0	0.0005513	0.0009694	0	0.0001378	4.51e-05	0	0.93e-05	0.0102877	0	0.0001224	0.0073641	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224
г	0.0006183	0	0	0	0	0.0004049	0	0.0005513	0.0009694	0	0.0001378	4.51e-05	0	0.93e-05	0.0102877	0	0.0001224	0.0073641	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	0.0001224	
д	0.00014	0.0001224	0.0009994	0.0001314	7.09e-05	0.002491	0.0000903	0.0009501	0.0029532	0.0003478	0	0.000430	0	0.0001515	0.000032	0	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	
е	0.0000123	0.0001224	0.0009994	0.0001314	7.09e-05	0.002491	0.0000903	0.0009501	0.0029532	0.0003478	0	0.000430	0	0.0001515	0.000032	0	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	0.0001344	
ж	0.0001675	1.29e-05	0	0	0.0001546	0.0006186	1.93e-05	0.0001083	0.0001083	0.0001083	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
з	0.000552	0.0001224	0.0004981	0	0.0001404	0	0.0001452	0.0001385	0.0001385	0.0001385	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
и	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
й	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
к	0.0001556	0.0001256	0.0001997	7.41e-05	0.0003671	0.0001174	0.0004026	7.73e-05	0.0029951	0.0004081	5.47e-05	0.0004734	0.0000951	0.0026408	0.0002183	0.0007745	0.0004161	0.0012609	0.0009758	0	0	0.0001158	0.0004281	0.0001701	0	0	0.0002222	0.0001256	1.59e-05	1.29e-05	0.000451
л	0.0001556	0.0001256	0.0001997	7.41e-05	0.0003671	0.0001174	0.0004026	7.73e-05	0.0029951	0.0004081	5.47e-05	0.0004734	0.0000951	0.0026408	0.0002183	0.0007745	0.0004161	0.0012609	0.0009758	0	0	0.0001158	0.0004281	0.0001701	0	0	0.0002222	0.0001256	1.59e-05	1.29e-05	0.000451
м	0.0004662	2.50e-05	0.0001159	0	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	0.0001159	
н	0.001607	0.0002029	0.0012399	0.0001256	0.0024701	0.0005491	0.0007697	0.0002666	0.0004301	0.0001051	0.0003183	0.000240	0.0007697	0.0002666	0.0004301	0.0001051	0.0003183	0.000240	0.0007697	0.0002666	0.0004301	0.0001051	0.0003183	0.000240	0.0007697	0.0002666	0.0004301	0.0001051	0.0003183	0.000240	
о	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
п	0.0012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
р	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
с	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
т	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
у	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
ф	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
х	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
ц	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
ч	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
ш	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
щ	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
ь	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
э	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	
я	0.00012302	0	0.0001552	0.0000000	0.0000748	0.0003643	0.0001452	0.0001452	0.0001452	0.0001452	0	0.000240	0	0.0001675	0.000032	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	0.0001675	

	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	4.51e-05	0.000955	0.007652	0.001397	0.001028	4.51e-05	0.002157	0.0002929	0.0001578	1.29e-05	0.0116969	0.0077421	0.0049001	0.0166469	9.02e-05	0.0018035	0.0114715	0.0023445	0.0102348	4.51e-05	0.0001675	0.000657	0.00076	0.0023864	0.0015233	0.0004026	0	6.4e-06	0	6.4e-06	0	
б	0.000628	0.0001224	2.58e-05	0	8.37e-05	0.002119	1.29e-05	0.0002921	0.0014235	1.7e-05	0	2.25e-05	9.34e-05	0	0.0047557	0	9.36e-05	0.0001224	9.66e-05	0.0000004	0	0	0	0	0	0	0.0002576	2.58e-05	0	0.0005153	1.93e-05	
в	0.0048823	6.4e-06	4.51e-05	0	0.0000921	0.0028321	0	0.0007958	0.0029693	0	0.0001578	4.51e-05	0	1.93e-05	0.0182767	0	0.0004009	0.0016264	0.0018325	0.0004638	0	0.0001578	1.93e-05	0	5.15e-05	0	0.0015104	3.86e-05	3.86e-05	0	0.0001932	
г	0.0006183	0	0	0	6.4e-06	0.0049807	0	0.0005379	0.0009694	0	0	7.73e-05	0	0.0003124	0.0073041	0	9.02e-05	2.55e-05	6.4e-06	0.0016586	0	0	0	0	0	0	0.0001546	0.0001803	0	5.15e-05	3.86e-05	
д	0.00314	3.22e-05	0.0009984	0.005134	7.09e-05	0.0024991	0.0006699	0.0005951	0.0025532	0.0003478	0	0.000438	0	0.0005185	0.000932	0	0.0003114	0.0001639	0.0001449	0.0034524	0	0	1.29e-05	0	0	0	0.0001256	0.0001642	0.00057	0.0003768	0.0005958	
е	0.0035232	0.0026247	0.0057647	0.0003156	0.0067645	0.002322	0.0049789	0.0003575	0.0019645	1.93e-05	0.0010145	0.004995	0.0054073	0.0176356	0.002892	0.0024766	0.0064636	0.0072075	0.0079225	0.0003349	0.0001707	6.44e-05	0.0009565	0.0064185	0.0031175	0.0028225	0.0010724	0.0008373	0	0.0001385		
ж	0.0018679	1.29e-05	0	0	0.0001546	0.0005668	1.93e-05	0.0001803	0.0007021	0	0	0.000248	0	6.4e-06	0.0027213	0	0.0002673	1.93e-05	0	0.0021803	0	0	0	0	0	0	4.51e-05	0	1.93e-05	4.83e-05		
з	0.0055522	3.22e-05	0.0004058	0	0	0.0014074	0	0.0001385	0.0018132	0	5.8e-05	1.93e-05	0	4.51e-05	0.001475	0	0.0002126	1.93e-05	0	0.0001897	0	0	0	0	0	0	5.15e-05	0.0001803	1.93e-05	1.29e-05	0.0001868	
и	7.73e-05	0.0015523	0.0040031	0.0007439	0.0036085	5.15e-05	0.0014525	0.0004813	0.0033882	0	0.0047277	0.018975	0.0028051	0.0101543	0.0010467	0.0015587	0.0006953	0.0038885	0.0047921	1.93e-05	0.00038	0.0003349	0.0002351	0.0026795	0.0019516	0.0006763	6.4e-06	3.22e-05	0	6.4e-06	3.86e-05	
й	0.0012302	0	0	0	0.0003226	0	0	0.0001821	0	0	0	0	0	0.0048887	0	0	0	0	0	0.0002319	0	0	0	0	0	0	0.0023891	0	3.22e-05	3.86e-05	1.29e-05	
к	0.0093556	0.0001256	0.0001997	7.41e-05	0.0003671	0.0012174	0.0004026	7.73e-05	0.0029951	0.0001481	5.47e-05	0.0004734	0.0001965	0.0009533	0.0026488	0.0002383	0.000715	0.004615	0.001269	0.0009758	0	0	0.0002158	0.0004283	0.0017101	0	0.0002222	0.0021159	5.15e-05	1.29e-05	0.000541	
л	0.013965	0.0008786	0.0006866	0.0018209	0.0016469	0.0008981	3.86e-05	0.0002641	0.0001672	0.0011159	0.0045467	0.0011787	7.73e-05	0	0.007735	0.0018724	0.0001159	0.0041986	0.0002361	0.0014364	2.58e-05	9.02e-05	0	6.4e-06	0.0003243	0	0.0045807	0	5.15e-05	2.58e-05	0.0015806	
м	0.0048662	2.58e-05	0.0001159	0	0.05e-05	0.0051625	6.4e-06	0.0001288	0.0028791	0.0001224	0	3.2e-06	2.58e-05	0	0.0001576	0	0.000438	0.0013558	3.86e-05	0.0021803	6.4e-06	1.29e-05	0	3.2e-06	1.29e-05	0	0.0007472	0.0003446	1.29e-05	7.73e-05	0.0004895	
н	0.003687	0.0002029	0.0012399	0.0021258	0.0024701	0.0091495	0.0007697	0.0026666	0.0046636	0.0010531	0.0006538	0.0037648	0.0026482	0.0114086	0.0037648	0.0026482	7.41e-05	0.0017487	0.0014086	0.003446	0.0002834	0	0.0002319	0	0.0014492	0.0007182	1.93e-05	0.0002512	0.001269	0.0002576	6.44e-05	0.0006763
о	6.44e-05	0.00257	0.0101628	0.0116293	0.0058015	0.0001675	0.0001997	0.0004605	1.000409	0	0.012314	0.0009434	0.0058019	0.016888	0.0006248	0.0127339	0.0056005	0.0038981	0.0221131	1.51e-05	0.0002866	0.0041899	0.0002973	0.0004187	0.0003897	6.4e-06	0	1.29e-05	0	0.0002861	0	
п	0.0013204	0	0.0001192	1.93e-05	3.54e-05	0.0000695	0	9.74e-06	0.0006312	2.58e-05	0	1.93e-05	0.000132	1.29e-05	0.0014041	3.86e-05	0.0001095	0.0026534	0.0001353	0.0006248	0	0	0	0	2.58e-05	0	0.0002287	1.29e-05	0	6.44e-05		
р	0.0032173	0.0013462	0.0008889	0.0011497	0.0019549	0.0093943	0	0.0002544	0.0007439	0	0.0024122	0	5.8e-05	0.0005668	0.0003927	0.0081769	7.41e-05	0.0005346	0.0042081	0.0011916	0.0001288	0.0001863	0	0	0	0	0.0003639	0	0.0001514	3.54e-05	6.4e-06	
с	0.0055217	4.51e-05	0.0057325	1.93e-05	0.0003114	0.0056243	1.29e-05	0	0.0009007	0.0009007	0.002235	0.0001288	0.0010112	0.0075719	0.000161	0.0002222	0.0003934	0.0024959	0.0015426	0	0	0	0	0	0	0.0005604	0.0014041	1.29e-05	0.0005339	0.0001061		
т	0.0078259	0	0.0004863	0.0001831	0.0003639	0.0106245	1.29e-05	3.86e-05	0.0072204	0.0011336	0.0012367	0.000132	9.02e-05	0.0006763	0.0095617	9.02e-05	0.0019033	0.0125117	0.0004187	0.0038531	0.0001353	7.41e-05	0	0.0064797	0.0001868	0	0.0007439	0.0001997	0.0038711	0.001819	0.0028115	
у	0.0001481	0.0019838	0.0006924	0.0008277	0.0029758	0.05e-05	0.0007955	0.0007665	0	0.0033976	0.001325	0.0028115	0.0026634	0.0001868	0.0008582	0.0035877	0.0000763	0.0025249	1.29e-05	0.0001831	0.0002287	0.000351	0.0012236	0.0003349	5.8e-05	0	0	0	0	0		
ф	0.0001353	0	0	6.4e-06	6.4e-06	0.0002319	0	0	0.0001771	0	0	0.0001831	0	1.29e-05	0.0002126	0	3.22e-05	1.93e-05	0	3.86e-05	1.29e-05	0	0	0	0	0	0	0.0001895	0	0	0	
х	0.0009178	3.54e-05	1.93e-05	0	0	5.47e-05	0.0000469	0	0	0.00219	0	6.4e-06	0	0	3.86e-05	0.0008389	0	0.0001127	0.0001578	1.29e-05	0.0003221	0	0	0	0	0	0	0.0009243	0	3.86e-05	0.0001095	
ц	0.0002931	6.4e-06	1.93e-05	0	0	0.0003382	0.0005282	6.4e-06	0	0.0009629	6.44e-06	6.4e-06	0	0	0.0006699	9.66e-05	0	6.44e-05	3.54e-05	9.02e-05	3.86e-05	0	0	0	0	0	0	6.4e-06	5.47e-05	0	3.22e-05	
ч	0.0011787	1.29e-05	0.0003885	6.4e-06	1.93e-05	0.0016562	6.4e-06	1.29e-05	0.0016167	0.0002126	0	0.0001997	3.22e-05	0.0002641	0.0033719	6.4e-06	0.0002093	0.0004122	0.000161	0.0014235	1.93e-05	0	0	0	0	0	0.0001288	0.0002892	0	0.0001417	0.0002222	
ш	0.0013776	0	6.12e-05	0	0.0001031	0.0020354	0	6.4e-06	0.001343	3.22e-05	0	4.51e-05	6.4e-06	1.29e-05	0.0019774	1.29e-05	0.0003478	0.0001159	0	0.0014525	2.58e-05	6.4e-06	0.0002061	0	0	0.0006763	0.0006763	3.22e-05	6.4e-06	0		
щ	0.0001226	0.0007761	1.93e-05	0	0	0.0010692	0	0	0.0001546	0	0	0	0	9.02e-05	0.0001578	0	1.61e-05	0	0	0	9.50e-05	0	0	0	0	0	0	0	0.0001353	0.0003221		
ы	0	0.0070046	0.0031046	0	0.0005571	0	0	0.0005791	0	0	0.001417	0.0016544	0	0.000161	0.0014106	0.0000022	0.0024057	0	1.93e-05	0	0.0002126	0	0	0	0	0	0	0.0004831	0.0014045	3.86e-05	0	0
ю	0	0.0001578	0.000019	0	0.0012463	0	2.58e-05	0.0002061	0	0	0.0050916	0.000132	0.0023859	0	0.0002415	0.0004031	0.0037422	0.0112074	0	0	0	0	0	0	0	0	0	0	0	0	0	
э	0	1.29e-05	0	0	1.29e-05	0	0	0	2.58e-05	0	0	0	0	0	1.93e-05	0	0	0	0.0005571	0.0001224	6.4e-06	1.29e-05	0	0	0	0	0	0	0	0	0	0
ю	0.0026666	1.29e-05	1.29e-05	0	3.86e-05	0.0001578	0	0	0.0003188	0	0	0.0012399	1.29e-05	0.0004348	6.4e-06	0.0011433	0.0003736	4.19e-05	0.001256	0	0	0	6.4e-06	0	0	0	0.0007633	0	2.58e-05	0.0002029		
я	0.0021807	0.0018885	0.0006261	0	0.00021705	8.37e-05	0	0.0005346	0.0008566	0	0	0.0013816	0.0005585	0.0031204	0.0010434	0.0006659	0.0015909	0.00071405	0.0000475	1.29e-05	0	0	0	0	0	0	1.93e-05	0.0003929	0	1.61e-05		

Таблиця частот символів:

Frequency with spaces:

'й': 0.1784963248350861, 'щ': 0.09482639906316687, 'б': 0.07109702041919348, 'ф': 0.0687922121086304, 'ц': 0.05665314084491001, 'к': 0.05447672257500722, 'з': 0.05056910956329784, 'г': 0.041341593092615814, 'ж': 0.04126290242253369, 'п': 0.03389082912010371, 'ю': 0.03242883930226225, 'р': 0.029912808666741907, 'э': 0.025143739898344078, 'ь': 0.024992570979502117, 'ш': 0.02438582449702684, 'ы': 0.01962089734284386, 'у': 0.01867868010896587, 'я': 0.017887631793929843, 'т': 0.014711013691141191, 'ч': 0.014578482036266044, 'х': 0.014419029888994385, 'е': 0.01415189577213667, 'д': 0.011807742126532527, 'о': 0.008927249440105528, 'н': 0.008291511658126314, 'м': 0.007740676967551488, 'а': 0.006197925672520493, 'в': 0.0059742785049186845, ' ': 0.003267733615515315, 'и': 0.0023130915389927803, 'л': 0.002081161142961275, 'с': 0.0010809613100754083

Frequency with NO spaces

'й': 0.11543027977827465, 'щ': 0.08654498156069845, 'б': 0.08373938448275428, 'ф': 0.06896273572182997, 'ц': 0.06631342527356462, 'к': 0.06155676607672665, 'з': 0.05032429475658357, 'г': 0.0502285061771103, 'ж': 0.04125462873102819, 'п': 0.03947497775436279, 'ю': 0.03641226396307602, 'р': 0.03060697189613493, 'э': 0.0

$$H(x_1, x_2, \dots, x_n) = - \sum_{z_1, z_2, \dots, z_n} P(x_1 = z_1, \dots, x_n = z_n) \cdot \log_2 P(x_1 = z_1, \dots, x_n = z_n).$$

Ми проходили циклом по значенням частоти кожного символу(біграми) та домножували на логарифм основи двійки по ній - загальну суму з мінусом отримували як результат ентропії (для n-грами потрібно ще ділити на n)

Оцінки для (10) Н , (20) Н , (30) Н (включно із відповідними скріншотами)

Оцінки для $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ умовних ентропій бралися з програми CoolPinkProgram шляхом 50+ експериментів з вгадування наступної літери в рядку

The screenshot displays the following information:

- Произвольная часть текста:** ло_иными_
- Использованные буквы:** ч, в,
- Порядок n-граммы:** A list from 5 to 50 symbols, with "10 символов" selected.
- Введенный символ:** в
- Символ по счету:** 2
- Номер эксперимента:** 55
- Поле ввода символов:** An empty input field.
- Buttons:** "Продолжить" (Continue) and "Другой" (Other).
- Неравенство для энтропии:** $2.97414488795797 < H < 3.77954878919313$
- Двоичная таблица угаданных символов:** A grid of 0s and 1s representing probability estimates for each symbol.
- Вероятности:** A vertical list of probabilities $q[1]$ through $q[32]$, where most values are approximately 0.037037 or 0.018518, except for $q[1]$ which is 0.333333.
- Строка состояния:** Вывести не удалось. Введите другую букву (Output failed. Enter another letter).

$$2.974 < H^{(10)} < 3.779$$

Нижні межі: 2.974, 1.800, 2.287

Отже верхня межа $H1_{\infty} = 2.479$, а нижня $H2_{\infty} = 1.800$

Оцінку надлишковості R російської мови у різних моделях відкритого тексту

За даними значеннями меж використовуючи формулу $R = 1 - \frac{H_{\infty}}{H_0}$ ми розраховуємо надлишковість російської мови. H_{∞} в даному випадку це ліміт куди прямує значення ентропії джерела символів. В наших експериментах $H_{\infty} = 3.923236$, тому $R = 0.2153528$

Межі зазначеної ентропії по експериментам програми CoolPinkProgram при $H1_{\infty} =$, $H2_{\infty} =$ межі $0.50 < R < 0.63$ (Значення такі низькі через відносно невелику кількість експериментів та людський фактор у виконанні завдання)

Висновки

В даній практичній роботі визначили ентропію російської мови на основі російського тексту через python. Обчислювалися частоти символів/біграм для розрахунку питомої ентропії на символ/біграму, визначення коефіцієнту хаотичності мови. Розібралися з переробкою письмових екземплярів в об'єкти аналізу частоти/ентропії. Ознайомилися з методикою програми CoolPinkProgram, що використовує принцип підрахунку умовної ентропії - як шанс вгадування людини наступних символів в контексті. Проаналізували наші труднощі й методи боротьби з ними. За нашим аналізом вийшло ± 0.215 і це значення відрізняється лише через людський фактор та не найбільший розмір текстового семплу мови.