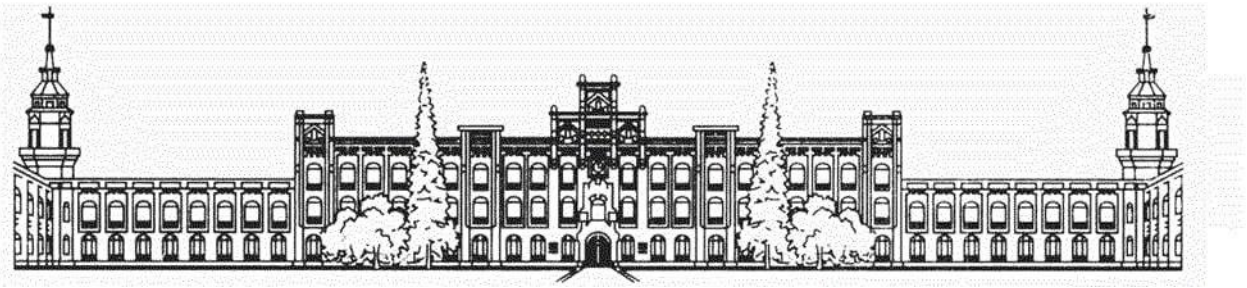


МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. ІГОРЯ СІКОРСЬКОГО”

КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали роботу:

студент ФБ-23 Хоменко Гліб

студент ФБ-23 Ткачук Андрій

Київ 2024

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:

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, атакож значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 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

1.

Частота літер тексту

Літера	Кількість	Частота			
о	130005	0,11012	**ы**	23093	0,01956
е	99389	0,08419	**ь**	22398	0,01897
а	93308	0,07904	**г**	20839	0,01765
т	78057	0,06612	**ч**	20530	0,01739
и	77636	0,06576	**з**	19878	0,01684
н	74310	0,06295	**б**	19275	0,01633
с	65232	0,05526	**й**	12356	0,01047
л	53670	0,04546	**ж**	11824	0,01002
р	50767	0,043	**х**	11164	0,00946
в	48436	0,04103	**ш**	10756	0,00911
к	45836	0,03883	**ю**	8088	0,00685
м	37707	0,03194	**щ**	4936	0,00418
у	34511	0,02923	**э**	3996	0,00338
п	33967	0,02877	**ц**	3269	0,00277
д	33947	0,02876	**ф**	2588	0,00219
я	28325	0,02399	**ъ**	444	0,00038
			ё	8	0,00001

AMOUNT_WITHOUT_SPACE.XLSX

Частота літер тексту(з пробілом)

Літера	Кількість	Частота
_	252492	0,17619
о	130005	0,09072
е	99389	0,06936
а	93308	0,06511
т	78057	0,05447
и	77636	0,05418
н	74310	0,05185
с	65232	0,04552
л	53670	0,03745
р	50767	0,03543
в	48436	0,0338
к	45836	0,03199
м	37707	0,02631
у	34511	0,02408
п	33967	0,0237
д	33947	0,02369
я	28325	0,01977
ы	23093	0,01611
ь	22398	0,01563
г	20839	0,01454
ч	20530	0,01433
з	19878	0,01387
б	19275	0,01345
й	12356	0,00862
ж	11824	0,00825
х	11164	0,00779
ш	10756	0,00751
ю	8088	0,00564
щ	4936	0,00344
э	3996	0,00279
ц	3269	0,00228
ф	2588	0,00181
ъ	444	0,00031
ё	8	0,00001

AMOUNT_WITH_SPACE.XLSX

Частота пересічних біграм тексту

	а	б	В	г	д	е	ж	з	и	й
а	0,00022	0,00168	0,00437	0,00109	0,00278	0,00437	0,00167	0,0042	0,00107	0,00073
б	0,00156	0,00002	0,00004	0,00002	0,00002	0,00219	0,00001	0,00002	0,00106	
в	0,00697	0,00015	0,00034	0,00025	0,00069	0,00492	0,00007	0,00067	0,0032	
г	0,00145	0,00005	0,00009	0,00001	0,00144	0,00017	0	0,00004	0,00103	
д	0,00465	0,00011	0,00123	0,00007	0,0001	0,00477	0,00009	0,00005	0,00276	
е	0,00022	0,00248	0,00345	0,00445	0,00402	0,00193	0,00106	0,00237	0,0011	0,00193
ж	0,0016	0,00001	0,00002	0,00003	0,00093	0,00363	0,00001	0,00001	0,00173	
з	0,00625	0,0002	0,00087	0,00044	0,00109	0,00038	0,00016	0,00014	0,00045	
и	0,00038	0,00113	0,00577	0,00109	0,0028	0,00238	0,00053	0,00287	0,00167	0,00128
й	0,00014	0,00024	0,00071	0,00023	0,00055	0,00012	0,00011	0,00025	0,00058	

BIGRAMM_CROSS_TABLE.XLSX

Частота пересічних біграм тексту(з пробілом)

	_	а	б	в	г	д	е	ж	з	и
_	0,01542	0,00177	0,0052	0,01379	0,00351	0,00667	0,00346	0,00135	0,00479	0,00946
а	0,01346	0,00002	0,001	0,00253	0,00058	0,00175	0,00329	0,0013	0,00311	0,00018
б	0,00021	0,00128	0,00001	0,00003	0,00002	0,00001	0,0018	0,00001	0,00001	0,00086
в	0,00533	0,00569	0,00001	0,00003	0,00001	0,00023	0,00396	0,00001	0,00044	0,00246
г	0,00058	0,00119	0	0,00002	0	0,00114	0,00013		0	0,00082
д	0,00124	0,00382	0,00007	0,00092	0,00001	0,00002	0,00391	0,00007	0,00001	0,0022
е	0,01761	0,00004	0,00129	0,0013	0,00331	0,00252	0,00132	0,00071	0,00119	0,00007
ж	0,00019	0,00132	0,00001	0	0,00002	0,00076	0,00299	0,00001		0,00141
з	0,0013	0,00514	0,00013	0,00063	0,00033	0,00082	0,0003	0,00013	0,00009	0,00033
и	0,01744	0,00013	0,00039	0,00313	0,00054	0,00158	0,00159	0,00031	0,00178	0,00032

BIGRAM_CROSS_SPACE_TABLE.XLSX

Частота непересічних біграм тексту

	а	б	в	г	д	е	ж	з	и	й
а	0,00023	0,00167	0,00441	0,00114	0,00277	0,0043	0,00172	0,00425	0,00105	0,00075
б	0,00157	0,00002	0,00005	0,00003	0,00001	0,00227	0,00001	0,00003	0,00107	
в	0,00697	0,00012	0,00035	0,00024	0,00066	0,00496	0,00006	0,00072	0,00315	
г	0,00146	0,00006	0,00008	0,00001	0,00144	0,00016	0	0,00005	0,001	
д	0,00461	0,00012	0,00123	0,00007	0,00011	0,00481	0,0001	0,00005	0,00278	
е	0,00022	0,00243	0,00343	0,00445	0,00398	0,00193	0,00108	0,00234	0,00111	0,00194
ж	0,00163	0,00002	0,00002	0,00003	0,00087	0,00364	0,00001	0,00001	0,00171	
з	0,00628	0,00022	0,00092	0,00043	0,0011	0,00034	0,00016	0,00013	0,00044	
и	0,00036	0,00115	0,00565	0,00102	0,00285	0,00238	0,00052	0,0028	0,00167	0,00132
й	0,00014	0,00023	0,00069	0,00023	0,00055	0,00011	0,00012	0,00022	0,00061	

BIGRAM_NO_CROSS_TABLE.XLSX

Частота непересічних біграм тексту(з пробілом)

	пробіл	а	б	в	г	д	е	ж	з	и
пробіл	0,01532	0,00175	0,00523	0,01364	0,00359	0,00663	0,00341	0,00132	0,00483	0,00943
а	0,01343	0,00002	0,00107	0,00254	0,0006	0,00173	0,00326	0,00124	0,00313	0,00018
б	0,00021	0,00129	0,00001	0,00002	0,00001	0,00001	0,00173	0,00001	0,00001	0,00085
в	0,00538	0,00564	0,00001	0,00004	0,00001	0,00022	0,00397	0,00001	0,00045	0,00246
г	0,00057	0,00122	0	0,00002	0	0,00111	0,00015		0	0,00083
д	0,00121	0,00384	0,00007	0,00093	0,00001	0,00002	0,00399	0,00007	0,00001	0,00221
е	0,01766	0,00004	0,00128	0,00129	0,00331	0,00248	0,00131	0,00067	0,00122	0,00006
ж	0,00016	0,00134	0,00001	0	0,00003	0,00081	0,003	0,00001		0,00143
з	0,00134	0,00509	0,00014	0,00064	0,00031	0,00084	0,0003	0,00011	0,00008	0,00031
и	0,01752	0,00011	0,00038	0,00308	0,00053	0,00161	0,00156	0,00029	0,00172	0,00029

BIGRAM_NOCROSS_SPACE_TABLE.XLSX

Значення ентропії

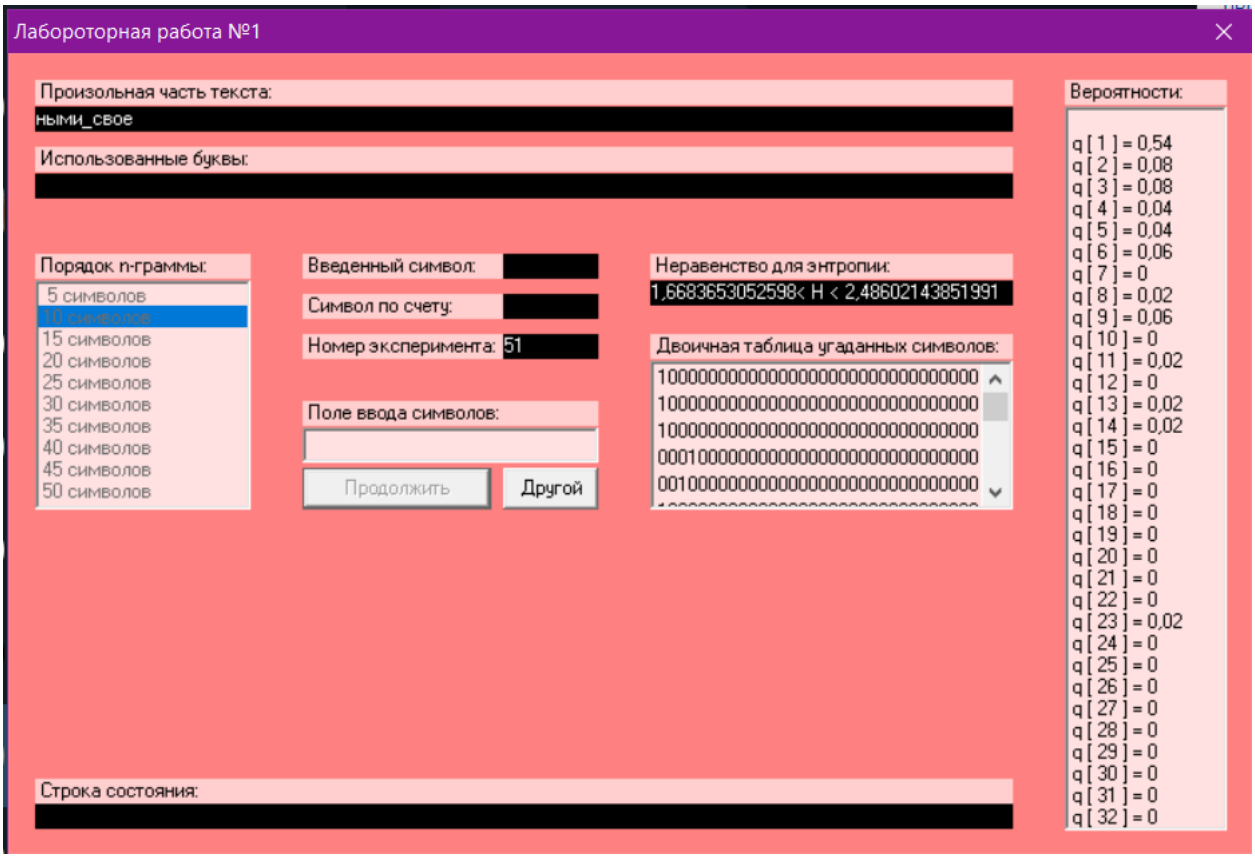
Позначення	Текст з пробілами	Текст без пробілів
H_1	4.360927156570439	4.478316752919476
H_2 з перетином	3.9805489517	4.162939067676659
H_2 без перетину	3.98083578	4.162153669819363

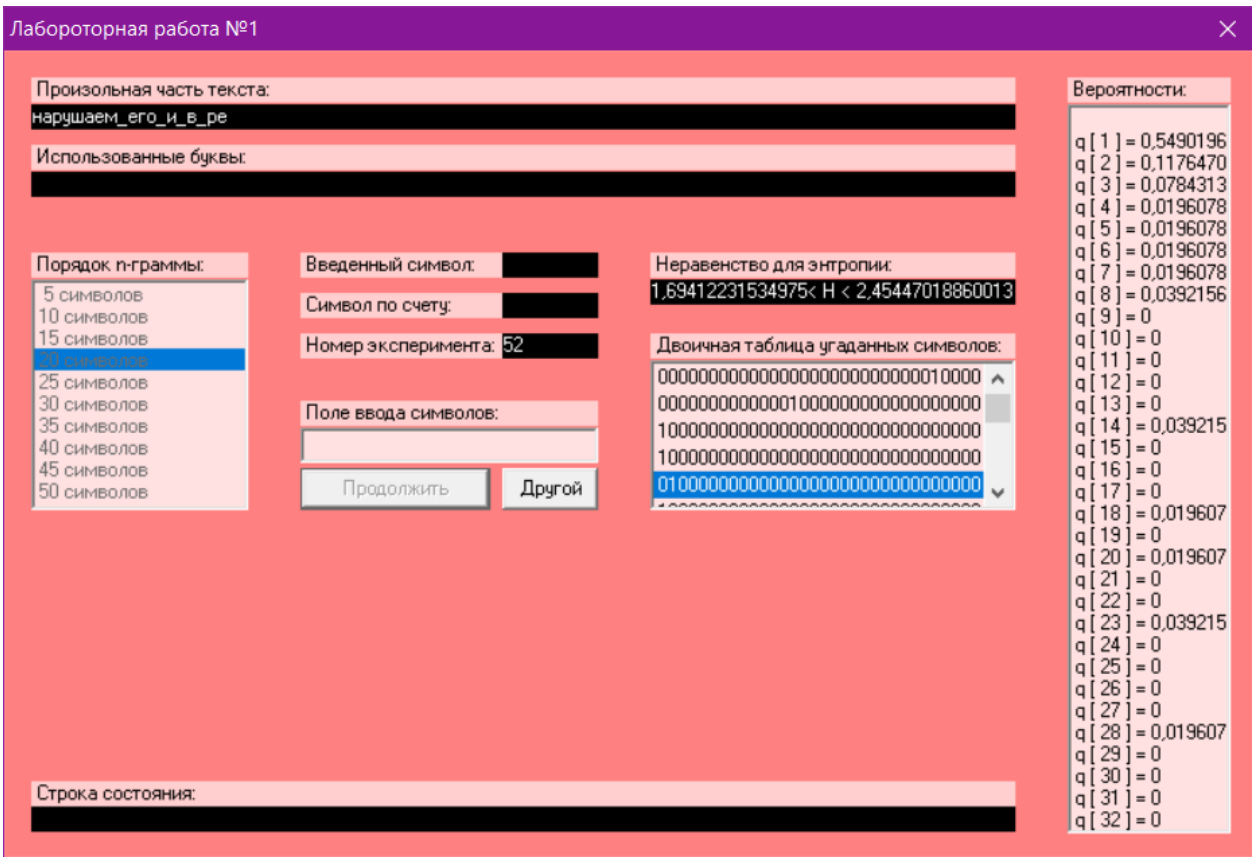
Оцінка надлишковості мови

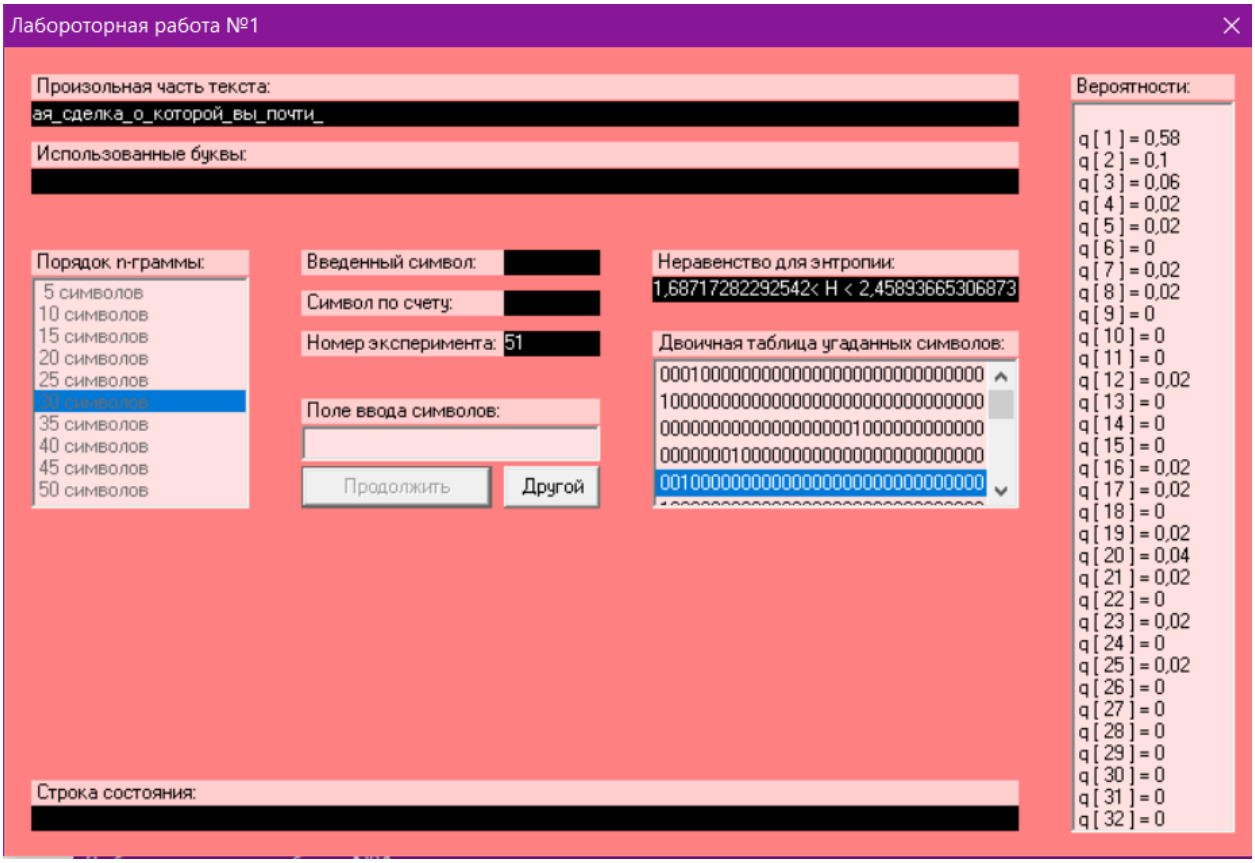
$$R = 1 - \frac{H_n}{H_0}$$

	Текст з пробілами	Текст без пробілів
H_1	0.14280904005607253	0.1122190996668142
H_2 з перетином	0.2175768008651957	0.17473952883639832
H_2 без перетину	0.21752042007341388	0.17489522600016316

2.







$H^{(10)}$:

$1{,}6683653052598 < H < 2{,}48602143851991$

$0{,}66632693894804 > R > 0{,}502795712296018$

$H^{(20)}$:

1,69412231534975 < H < 2,45447018860013

0,66117553693005 > R > 0,509105962279974

$H^{(30)}$:

1,6871728229254 < H < 2,45893665306873

0,66256543541492 > R > 0,5082126693862

Висновки:

Під час виконання комп'ютерного практикуму, ми навчились вимірювати частоту повторювання символів та біграм у тексті, визначати ентропію та надлишковість мови. За результатами роботи можна зазначити, що в тексті без пробілів ентропія більша, ніж в текстах з пробілами.