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 Дэвида 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 H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також 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 (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 H1 та H2

Таблиця Частоти Біграм:

\	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	28	612	4754	832	5032	28	1376	3908	98	8	7264	4808	3048	10338	56	1120	7124	1456	6356	28	104	408	472	1482	946	250	0	4	0	4	0
б	390	76	16	0	52	1378	8	182	884	54	0	14	58	0	2954	0	62	76	60	502	0	0	0	0	0	0	160	16	0	320	12
в	3032	4	28	0	554	1262	0	606	1844	0	98	28	0	12	6382	0	254	1010	1138	288	0	98	12	0	32	0	938	24	24	0	120
г	384	0	0	0	4	2918	0	344	602	0	0	48	0	194	4536	0	56	16	4	1030	0	0	0	0	0	0	96	112	0	32	24
д	1950	20	620	2244	44	1552	416	590	1834	216	0	272	0	322	3746	0	212	226	90	2144	0	0	8	0	0	0	78	102	354	234	370
е	2188	1630	3580	196	4766	1442	3092	222	1220	12	630	3102	3358	10952	1796	1538	4014	4476	4920	208	106	40	594	3986	1936	1256	666	520	0	0	86
ж	1160	8	0	0	96	352	12	112	436	0	0	154	0	4	1690	0	166	12	0	1306	0	0	0	0	0	0	28	0	0	12	30
з	3448	20	252	0	0	874	0	86	1126	0	36	12	0	28	916	0	132	12	0	242	0	0	0	0	0	0	32	112	12	8	116
и	48	964	2486	462	2236	32	902	300	210	0	2936	6812	1742	6306	650	968	4338	1918	2976	12	236	208	146	1664	1212	420	4	20	0	4	24
й	764	0	0	0	0	2436	0	0	672	0	0	0	0	0	3036	0	0	0	0	144	0	0	0	0	0	0	1434	0	20	24	8
к	5810	78	124	46	228	756	250	48	1860	92	34	294	122	592	1640	148	444	2866	824	606	0	0	134	266	1062	0	138	1314	32	8	336
л	8630	422	314	634	652	5588	24	164	5072	72	922	732	48	0	4816	666	72	2608	184	892	16	56	0	4	574	0	2800	0	32	16	932
м	3022	16	72	0	50	3206	4	80	1788	76	0	2	16	0	5066	0	272	842	24	1354	4	8	0	2	8	0	464	214	8	48	304
н	2240	126	770	134	1534	5682	478	1656	2772	654	406	154	2338	1272	7080	46	1086	1142	898	176	0	136	0	900	446	12	156	824	160	40	420
о	40	1596	6274	7222	3106	104	124	286	254	0	7658	5554	3418	9122	388	7908	6018	1924	14042	10	178	2602	166	260	242	4	0	8	0	128	0
п	820	0	74	12	22	540	0	6	392	16	0	12	82	8	872	24	68	1654	84	388	0	0	0	0	16	0	142	8	0	0	40
р	1998	836	552	714	1214	5834	0	158	462	0	1498	0	36	352	5212	5078	46	332	2658	740	80	66	0	0	0	0	226	0	94	22	4
с	3498	28	3560	12	212	3504	8	28	1858	378	188	1388	80	628	4706	100	138	580	1550	958	0	0	0	0	0	0	348	872	8	344	128
т	4860	0	302	64	226	6598	8	24	4484	704	768	82	56	420	5938	56	1182	7770	260	1896	84	46	0	4024	116	0	462	124	2404	734	1746
у	92	1232	430	514	1848	50	494	476	0	0	2110	1138	1746	1654	116	528	2228	420	1568	8	64	142	218	822	208	36	0	0	0	0	0
ф	84	0	0	4	4	144	0	0	110	0	0	64	0	8	132	0	20	12	0	24	8	0	0	0	0	0	0	68	0	0	0
х	570	22	12	0	34	398	0	0	1360	0	4	0	0	24	516	0	70	98	8	200	0	0	0	0	0	0	574	0	0	24	68
ц	182	4	12	0	210	328	4	0	598	40	4	0	0	416	60	0	40	22	56	24	0	0	0	0	0	0	4	34	0	0	20
ч	732	8	240	4	12	972	4	8	1004	132	0	124	20	164	2094	4	130	256	100	884	12	0	0	4	0	0	80	174	0	88	138
ш	1166	0	38	0	64	1264	0	4	834	20	0	28	4	8	1228	8	216	72	0	902	16	4	0	128	0	0	420	420	0	20	4
щ	132	482	12	0	0	664	0	0	96	0	0	0	0	56	98	0	10	0	0	62	0	0	0	0	0	0	50	4	0	84	200
ы	0	4350	1928	0	346	0	0	422	0	0	0	880	906	1642	0	100	876	374	1494	0	12	0	132	0	0	0	0	0	0	0	0
ь	0	98	118	0	774	0	16	128	0	0	0	3162	82	1432	0	150	300	2324	6960	0	0	0	0	300	1144	24	0	0	0	0	0
э	0	0	8	0	8	0	0	0	16	0	0	0	0	0	12	0	346	76	4	8	0	0	0	0	0	0	0	0	0	0	0
ю	1656	8	8	0	24	98	0	0	198	0	0	770	8	262	270	4	710	232	26	780	0	0	0	0	4	0	0	474	0	16	126
я	1358	676	164	0	168	52	0	332	550	0	0	858	404	2062	648	416	988	4186	340	8	0	0	0	0	0	0	12	244	0	0	10

Якщо хочете подивитися на результати інших біграм(без пробілів, без перетинів) використайте код!!!!

Таблиця частот символів:

'й': 172393, 'щ': 91584, 'б': 68666, 'ф': 66440, 'ц': 54716, 'к': 52614,

'з': 48840, 'г': 39928, 'ж': 39852, 'п': 32732, 'ю': 31320, 'р': 28890,

'э': 24284, 'ь': 24138, 'ш': 23552, 'ы': 18950, 'у': 18040, 'я': 17276,

'т': 14208, 'ч': 14080, 'х': 13926, 'е': 13668, 'д': 11404, 'о': 8622,

'н': 8008, 'м': 7476, 'а': 5986, 'в': 5770, ' ': 3156, 'и': 2234,

'л': 2010, 'с': 1044.

Розрахунок H1 H2 проводили за формулами

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

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)

Значення H1

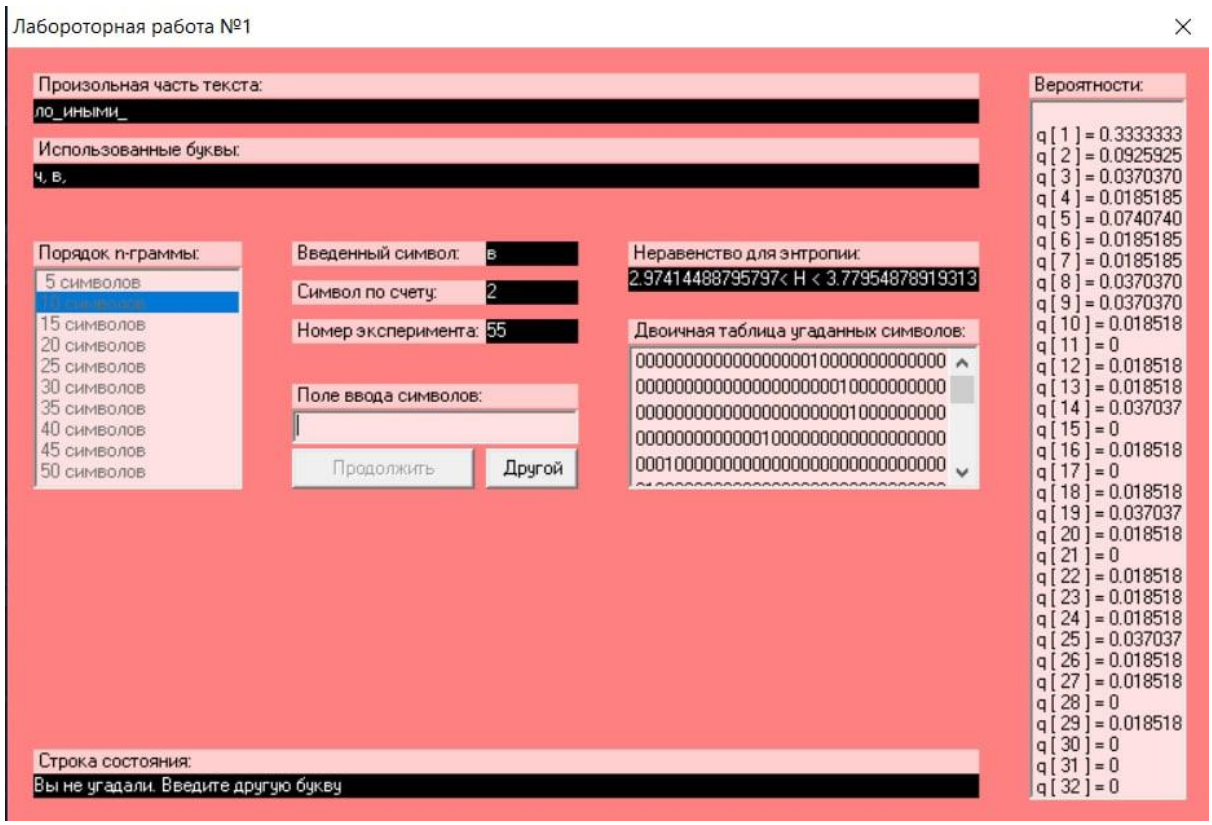
```
H1 (spaces): 4.328

H1 (NO spaces): 4.301
```

Значення H2

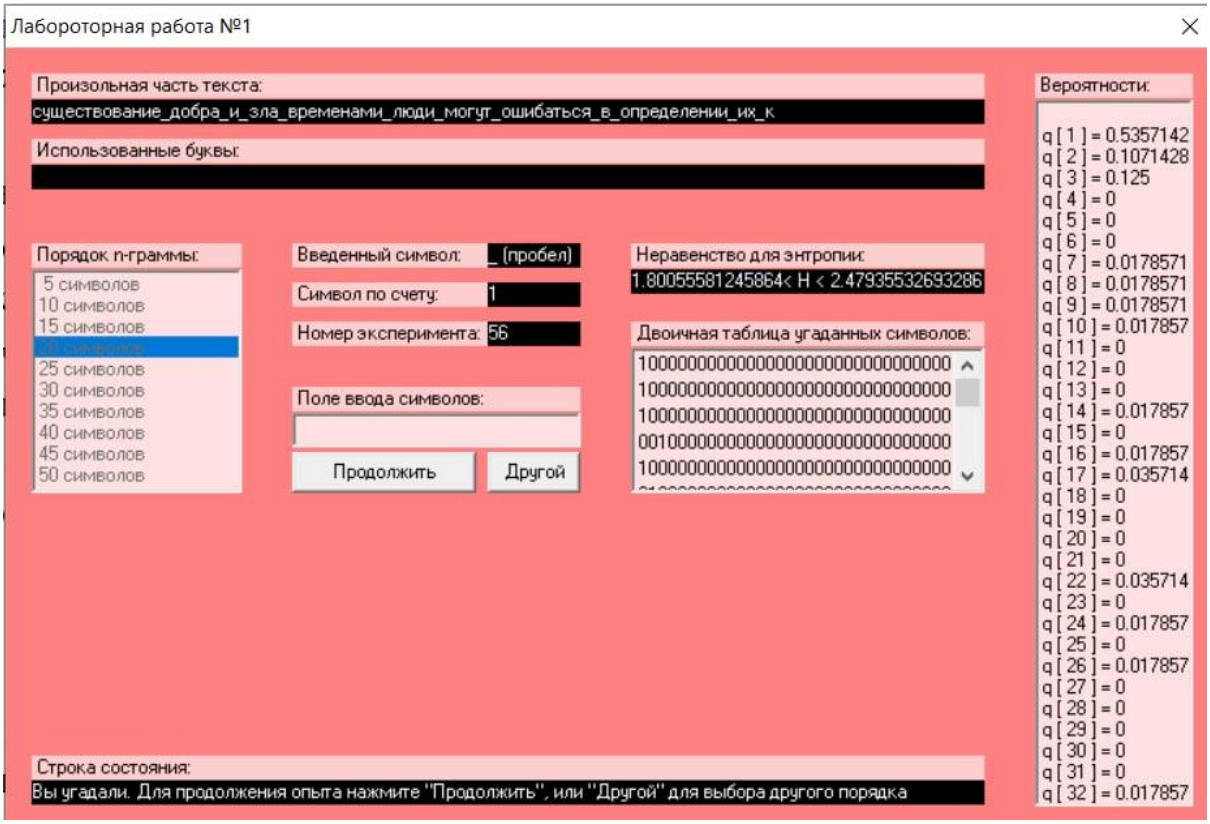
```
H2 (intersection, spaces): 3.924

H2 (NO intersection, spaces): 2.211

H2 (intersection, NO spaces): 2.730

H2 (NO intersection, NO spaces): 1.525
```

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

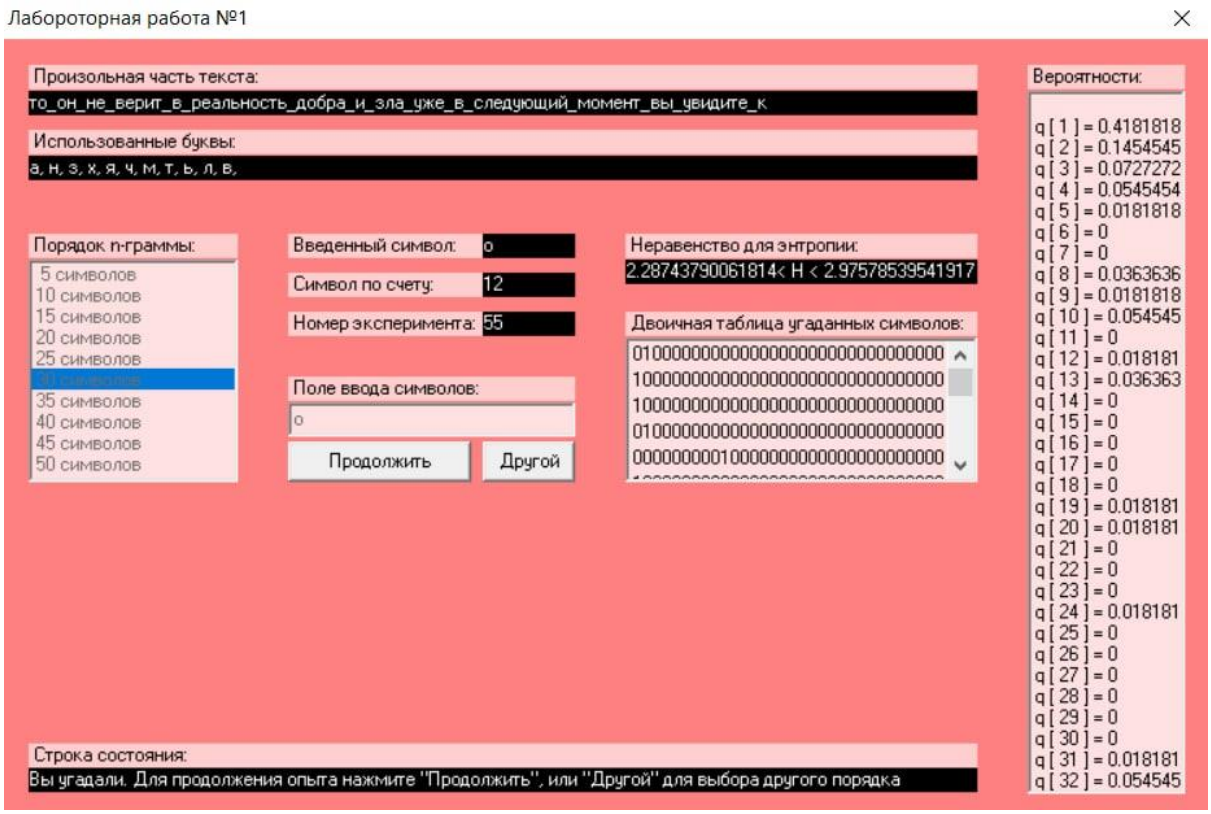
Оцінки для $H^{(10)}, H^{(20)}, H^{(30)}$ умовних ентропій бралися з програми CoolPinkProgram шляхом 50+ експериментів з вгадування наступної літери в рядку



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



$1.800 < H^{(20)} < 2.479$



$2.287 < H^{(30)} < 2.975$

Верхні межі для умовної ентропії - 3.779 , 2.479 , 2.975

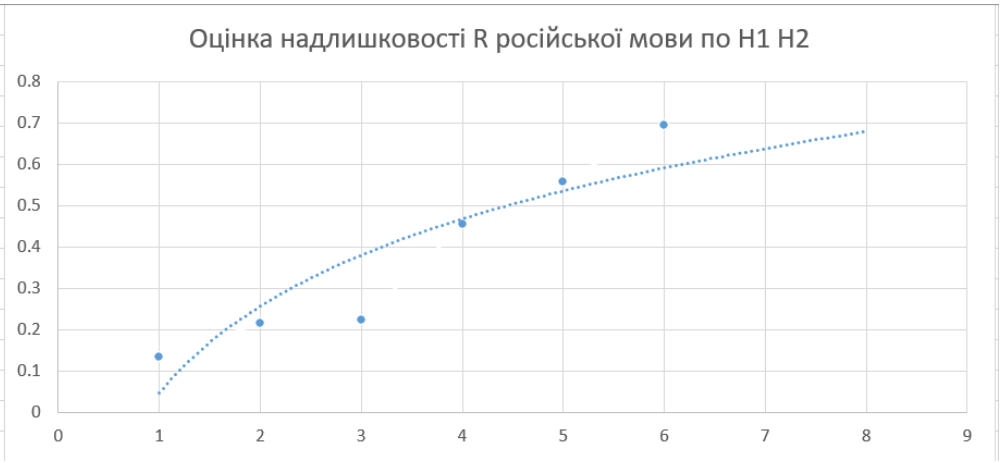
Нижні межі: 2.974, 1.800, 2.287

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

Оцінку надлишковості R російської мови у різних моделях відкритого тексту

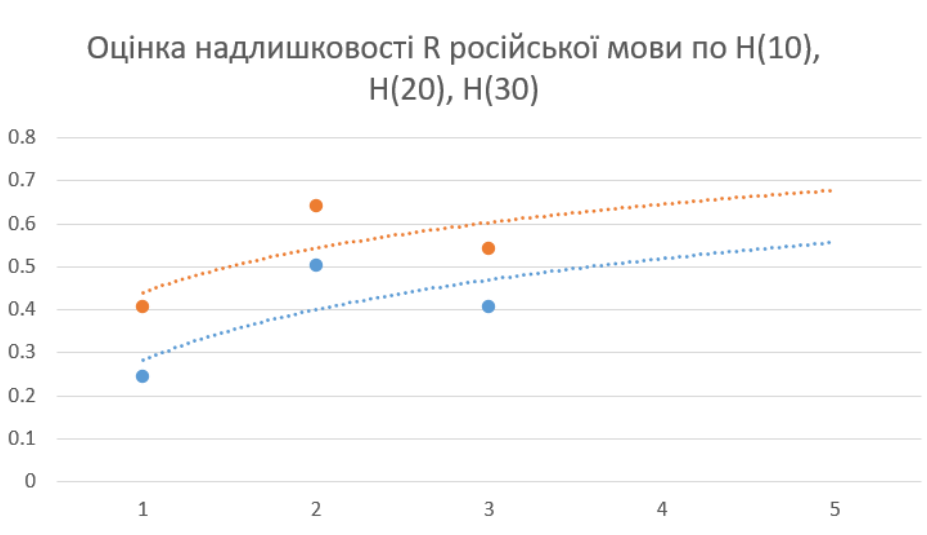
Діаграма для H1, H2 з нашого тексту

4.32932	0.13414
3.92475	0.21505
3.88557	0.22289
2.73136	0.45373
2.21188	0.55762
1.52577	0.69485



$H^{(10)}$Діаграма для H(10), H(20), H(30)

3.77955	0.24409
2.47936	0.50413
2.97578	0.40484
2.97414	0.40517
1.80056	0.63989
2.28744	0.54251



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_∞= 1.52 , тому R = 0.695

Межі зазначеної ентропії по експериментам програми CoolPinkProgram при $H1_\infty$= 2.479, $H2_\infty$= 1.800 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 мові 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 наши труднощі й методи боротьби з ними. Російська мова взагалі має ентропію 0.72, за нашим аналізом вийшло $\pm$ 0.695 і це значення відрізняється лише через людський фактор та не найбільший розмір текстового семплу мови.