Міністерство освіти і науки, молоді та спорту України
Національний технічний університет України
“Київський політехнічний інститут”
Фізико-Технічний інститут

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1
за семестровий курс предмету «Симетрична криптографія»

Роботу виконали:
Студенти групи ФІ-03
Піжук Богдан
Швець Катерина

Приймав:
Чорний Олег Миколайович

Київ-2023

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.

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

Хід роботи

- Очистка вхідного текстового файлу та редагування;
- Написання функцій для підрахунку частот літер та частот біграм. Визначили частоти та значення ентропії;
- Результати оформили у вигляді таблиць (див. нижче);
- Було проведено по 50 експериментів для умовних ентропій джерел;
- Обрахували оцінки надлишковості російської мови для даних умовних ентропій джерел;
- Оформлення звіту.

Таблиці частот букв і біграм тексту, одержані значення H1 та H2

	Текст з пробілами			Текст без пробілів		
	Літери	Біграми з ∩	Біграми без ∩	Літери	Біграми з ∩	Біграми без ∩
H	4.35676410 5666061	3.970778960 772621	3.9707789607 72621	4.51311878 1529859	4.191947557 3122496	4.1919438062 70274
R	0.12864717 886678778	0.205844207 84547575	0.2058442078 4547575	0.08903109 631167849	0.153859214 55645768	0.1538599717 008542

Текст з пробілами				Текст без пробілів			
Літера	Частота	Літера	Частота	Літера	Частота	Літера	Частота
_	0.1719224 18205326 42	я	0.0182418 18200896 088	о	0.1126614 90145071 01	ь	0.0196366 76946528 837
0	0.0951270 04279916 56	ь	0.0165804 50423036 7	е	0.0813193 01908298 66	г	0.0195883 49857745 76
е	0.0686628 72918772 53	г	0.0165396 44898668 226	а	0.0811667 26385140 66	ы	0.0189083 18679869 626
а	0.0685340 44048980 61	ы	0.0159654 52877197 524	и	0.0656689 19399390 94	б	0.0175427 33228256 436
и	0.0554482 95320072 706	б	0.0148124 05345756 867	н	0.0638435 36217355 92	з	0.0168792 71337963 07
н	0.0539070 12371069 12	з	0.0142522 03789783 929	т	0.0570439 14825577 184	ч	0.0140473 03935274 839
т	0.0481656 75092424 51	ч	0.0118610 00061791 223	с	0.0523009 56255009 62	й	0.0113665 31281779 375
с	0.0441609 04343689 78	й	0.0095974 59331465 606	л	0.0484506 68052963 726	ж	0.0100430 59436105 721
л	0.0409098 69923647 035	ж	0.0084799 70899831 765	в	0.0449869 96561182 44	ш	0.0091766 23772923 437
в	0.0379852 79698552	ш	0.0077483 86141511	р	0.0429047 89421614	х	0.0086892 10563196

	1		227		5		988
р	0.036227144534333185	х	0.007336833281452024	к	0.033200709993972 92	ю	0.006207269360694751
к	0.028033395241143 16	ю	0.005241178137099567	м	0.030441923611441 92	ц	0.003578275730895439
м	0.025703982735765577	ц	0.003021357611454 4605	д	0.029886852477419 17	э	0.003086720199273 2986
д	0.025235302141590508	э	0.002606307135020 8166	у	0.026790466860389255	щ	0.002939667771976226
у	0.022620833901695996	щ	0.002482141753728 1678	п	0.024891902658197	ф	0.001861283305131025
п	0.021017759730077287	ф	0.001571595624248741	я	0.021604279846982 63		

-Біграми (дуже багато, в таблиці наведено 5 найчастіших)

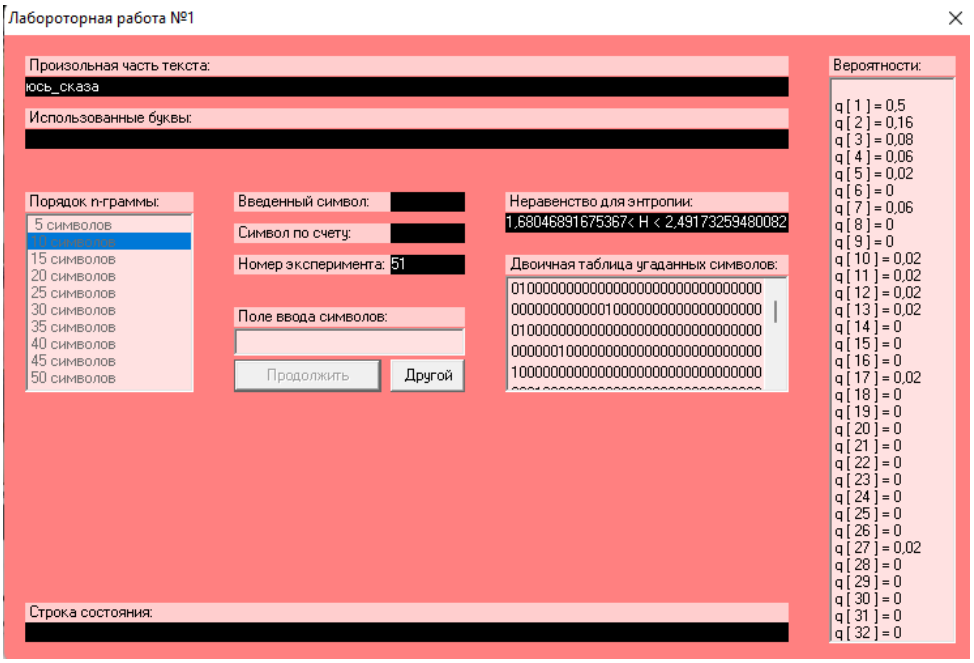
Текст з пробілами				Текст без пробілів			
з перетином		без перетину		з перетином		без перетину	
'о_'	0.021034094201356726	'о_'	0.02107896801663	'то'	0.016309713337319 17	'то'	0.016400844481940156
'и_'	0.018540875209055566	'и_'	0.018507054109291186	'ст'	0.013205731320531709	'ст'	0.013100792426725728
'а_'	0.016745431090213488	'а_'	0.016607848417969 82	'на'	0.012268875538329 62	'на'	0.012265423601033372
'е_'	0.016596782307647016	'е_'	0.016477270739990696	'ов'	0.011162184441152063	'ов'	0.011228461637240051
'_с'	0.016061646690407723	'_с'	0.015851197408965 79	'го'	0.010625062997855657	'но'	0.010572593550952665

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

- $H^{(10)}$

1,68046891675367<$H^{(10)}$<2,49173259480082

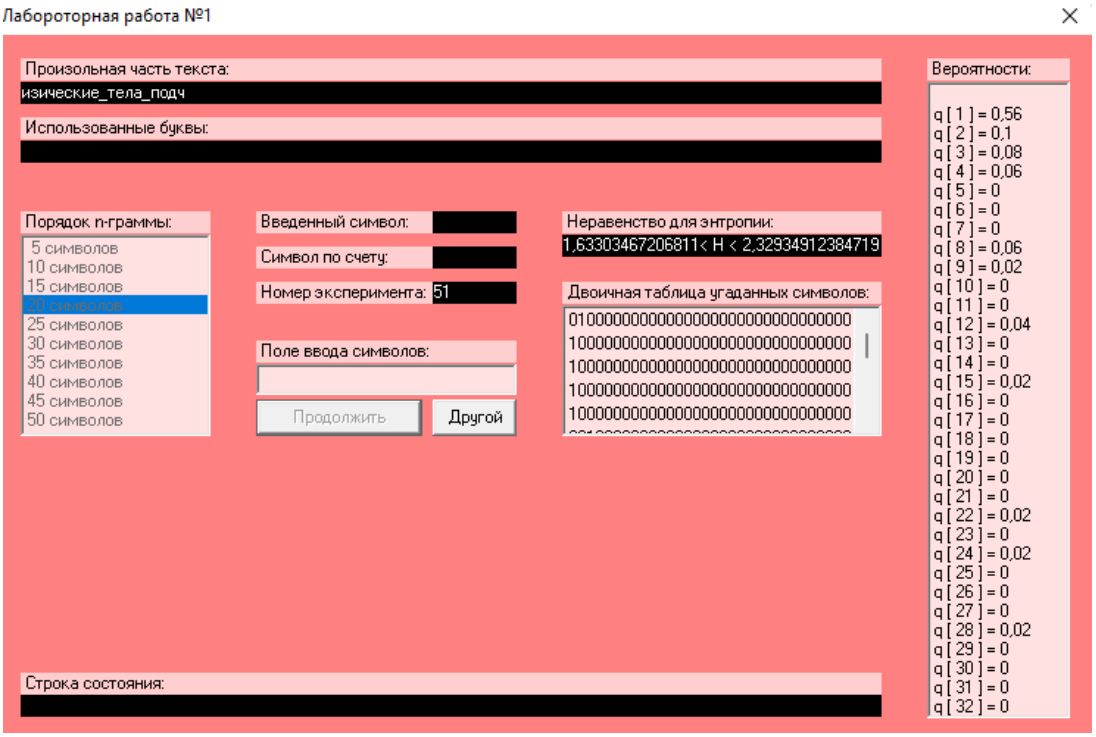
R=1 $-\frac{1,68046891675367+2,49173259480082}{2} * 1/log_2 32 = 0,58278$



- $H^{(20)}$

1,63303467206811<$H^{(20)}$<2,32934912384719

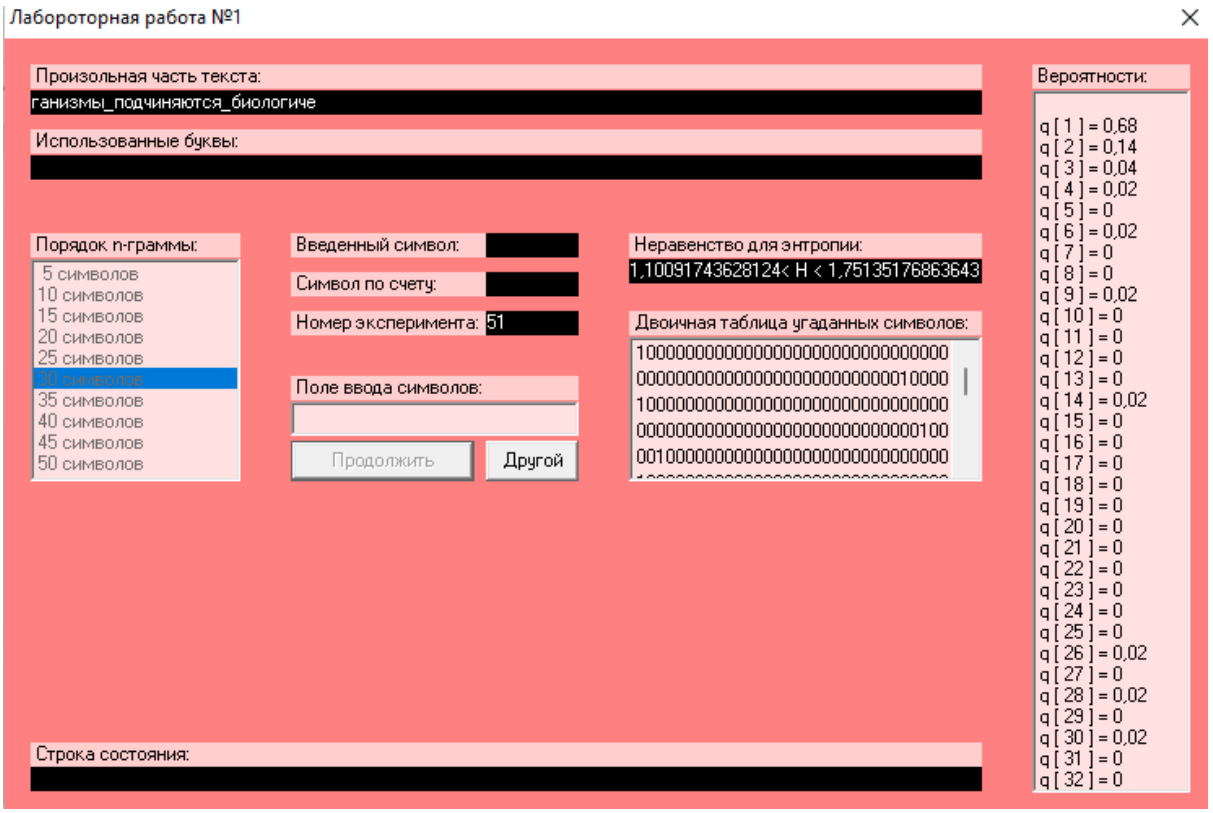
R=1 $-\frac{1,63303467206811+2,32934912384719}{2} * 1/log_2 32 = 0,603762$



- $H^{(30)}$

$1{,}10091743628124 < H^{(30)} < 1,751351176863643$

$R = 1 - \frac{1{,}10091743628124 + 1{,}751351176863643}{2} * 1/log_2 32 = 0,714773$



Висновки: відомо, що пробіл є найвживанішим символом в російській мові. В даному комп'ютерному практикумі ми розглянули над одним текстом два випадки (з пробілом та без пробілу) та порівняли оцінки ентропії та надлишковості в літерах та біграмах(з перетином та без). Переконалися на власному дослідженні, що найчастіший символ є " ", а найчастіші літери ("о", "е", "а"). Навчилися оцінювати ентропію та надлишковість на символ джерала.