Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

КРИПТОГРАФІЯ

Комп’ютерний практикум №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Роботу виконали:
Касаб О.Р.
Косигін О.С.
Групи ФБ-06

Київ – 2022

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, 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.

Хід Роботи
Для виконання даної роботи потрібно було обрати текст російською мовою. Ми обрали текст ранобе

“No game no life”, коротко – “NGNL”

Найбільшими труднощами особисто для нас стали розбір та зрозуміння завдань лабораторної роботи, підрахунок необхідних n-грам та вивід результатів

частоти нграм з пробілами і без

N-gramma with spaces				N-gramma without spaces			
ъ	0.00023589313677991353	ы	0.01733996011591426	ъ	0.0002774369097796511	ы	0.02039374699887958
ф	0.0025313148139075336	у	0.02164228801826176	ф	0.0029771114549431788	у	0.02545376940724537
щ	0.0029468496471583045	п	0.022589489690562643	щ	0.0034658272421704103	п	0.026567785306514432
ц	0.002955922460111378	д	0.023979444634973517	ц	0.003476497892546551	д	0.028202528944139146
э	0.0049319811212908075	м	0.026993433097984566	э	0.005800565544469936	м	0.03174731899909299
ю	0.005387436331535102	к	0.0274416300578664	ю	0.006336232193352185	к	0.03227444912767433
ш	0.006958847535007449	в	0.03684832252761311	ш	0.008184388838499706	в	0.04333777943765672
х	0.007205628047331051	р	0.039539318849494735	х	0.008474630528730726	р	0.046502694339219976
ж	0.008564735427701475	л	0.04139743094228421	ж	0.010073093955076561	л	0.048688043536253535
й	0.00902019063794577	с	0.04785545920228199	й	0.010608760603958812	с	0.05628341247399029
ч	0.012536812938557096	н	0.053074141212889926	ч	0.01474470468975084	н	0.06242117057034626
з	0.013975761072914568	т	0.0553024240741648	з	0.016437069839406713	т	0.06504188230272635
г	0.014603599729267261	и	0.059911413054326186	г	0.01717547884543563	и	0.07046257269380568
б	0.014603599729267261	е	0.0678392370127219	б	0.01717547884543563	е	0.07978658699247719
я	0.01640546038174768	а	0.07059555758786566	я	0.019294670010137117	а	0.08302833057674865
ь	0.01673389621064894	о	0.09831118659691489	ь	0.0196809475537534	о	0.11562503334578242
			0.14974133410270787				

частоти біграм

на скріншоті вказані дані з кінця таблиці, для перегляду повних таблиць значень зверніть увагу на **csv** файли прикріплені до лаби:

Block		Block without spaces		Slide		Slide without spaces	
ы	0.0058682847696779885	за	0.005702383391381547	в	0.005797527477014028	ис	0.005730139251987409
те	0.0059009468370416875	ис	0.005762138905961893	ы	0.005801156602195258	за	0.005789894894093795
м	0.005959012734577153	ил	0.005838967424708051	м	0.005810229415148332	ил	0.005915808568532252
го	0.006027965987900519	со	0.0059371371986614765	го	0.006098744867056072	со	0.005969161820412954
во	0.00611143571560775	ит	0.006090794236153794	ка	0.006211247747674184	ит	0.006058795283572534
ка	0.00613321042718355	ет	0.00611213549136106	ол	0.0062747574383457	од	0.006220989169289868
ть	0.006202163680506916	ом	0.006197500512190125	ть	0.006312863252748609	об	0.0063511711038787815
ол	0.006394506966093145	ло	0.006308475039267909	во	0.0063527836297421324	ло	0.00635330523395401
от	0.006419910796264911	ва	0.006381035306972615	та	0.006361856442695206	ом	0.006355439364029237
та	0.006419910796264911	об	0.006385303558014068	от	0.006396333131916886	ет	0.006385317185082431
д	0.006459831100820544	од	0.006415181315304241	й	0.006470730198132089	ва	0.006413060876060396
м	0.006477976693800376	ер	0.006419449566345694	д	0.006501577762172539	ре	0.006434402176812677
й	0.006514267879760042	ре	0.006517619340299119	м	0.006565087452844055	ер	0.006498426079069519
ос	0.0068481467905889696	те	0.007059687222563682	ос	0.006672146645690323	те	0.006878301232460119
ов	0.0069570203484679675	ат	0.007209076009014546	ов	0.0069080397824702365	ан	0.007115189670810436
ли	0.007015086246003433	го	0.0072346855152632655	пр	0.007062277602672488	го	0.007204823133970016
ен	0.007025973601791332 5	ан	0.0072944410298436116	ен	0.00707860866598802	ат	0.007283785946753454
пр	0.007080410380730832	ть	0.007328587038175237	ли	0.007134860106297077	ка	0.007324334418182788
ал	0.00715662187124613	ка	0.007473707573584648	ал	0.007154820294793839	ть	0.0074246385317185085
ор	0.0075739705097822895	ас	0.007495048828791914	ор	0.007470554185560799	та	0.007552686336232193
т	0.00773365172800482	та	0.007648705866284231	т	0.007795360889280834	ас	0.007595368937736755
ла	0.00790422030201525	во	0.007721266133988936	ни	0.008131054968544557	ол	0.007717014352024756
ни	0.008205437145480478	ак	0.007853581916273987	ла	0.008145571469269475	ак	0.007898415408419143
ко	0.008299794228975609	ол	0.007951751690227412	к	0.008183677283672385	во	0.007928293229472336
к	0.008390522193874773	ес	0.008203578501673154	ко	0.008452232547083363	ес	0.008263351651283146
не	0.0087788378836432	пр	0.008378576794372738	не	0.0086590926824134 4	пр	0.008310302512938163
ро	0.008891340560118165	ли	0.008583452844362493	ро	0.009114547892657735	ли	0.00867097049565171
о	0.009203444759371291	от	0.008660281363108653	о	0.0091472100192888	ал	0.008720055487381955
по	0.009217961233755158	ал	0.008664549614150106	ь	0.009237938148819536	от	0.008875846982873607
ь	0.009486516009856685	он	0.009121252475585604	по	0.009288745901356749	он	0.00901669956783866
я	0.009838540513665447	ор	0.009304787270368094	я	0.009756903049735347	ор	0.009334684949047644
и	0.00986757346243318	ла	0.009787099638052311	на	0.009891180681440835	ла	0.009614255988902524
на	0.009874831699625112	ен	0.009855391654715565	и	0.009956504934702966	ни	0.00975084031371712
ра	0.010281292982373371	ни	0.009868196407839924	ра	0.010232318448476403	ен	0.00991943658966014
но	0.010803886060192561	не	0.009953561428668989	но	0.010464582460075087	не	0.010226751320492984
ст	0.012132143466316336	ко	0.01015416922761729	ст	0.012094059666447105	ко	0.010265165661847089
н	0.013148296673186984	ос	0.01061087208905279	н	0.013239048661124993	ос	0.010585285173131303
в	0.013921298934127 87	ов	0.010751724373420746	в	0.013661841744738222	ов	0.01064290668516246
то	0.014759625329796152	по	0.010760260875503653	то	0.014746950173925825	ро	0.010747479058848637
п	0.01489027359925095	ро	0.010764529126545106	п	0.014854009366772093	по	0.010924611855092569
и	0.015877393857353865	на	0.011801714129618248	и	0.015915528482281702	на	0.011645947820519661
е	0.017060486519638974	ра	0.012074882196271257	с	0.017096808728771885	ра	0.012042896014512085
с	0.01716210184032604	но	0.012535853308748207	е	0.017118583479859263	но	0.012399295737075175
а	0.017423398379235636	ст	0.014204739465956429	а	0.017447019308760525	ст	0.014475804300272102
о	0.021625917713364955	то	0.01770897357098955	о	0.021591480265724546	то	0.017640719201835353

---H1 entropy---

H1:	4.401484367688899
H1 without spaces:	4.460162425027035
H1 with redundancy:	0.1274503412019915
H1 without spaces with redundancy:	0.11581801114416501

---H2 block entropy---

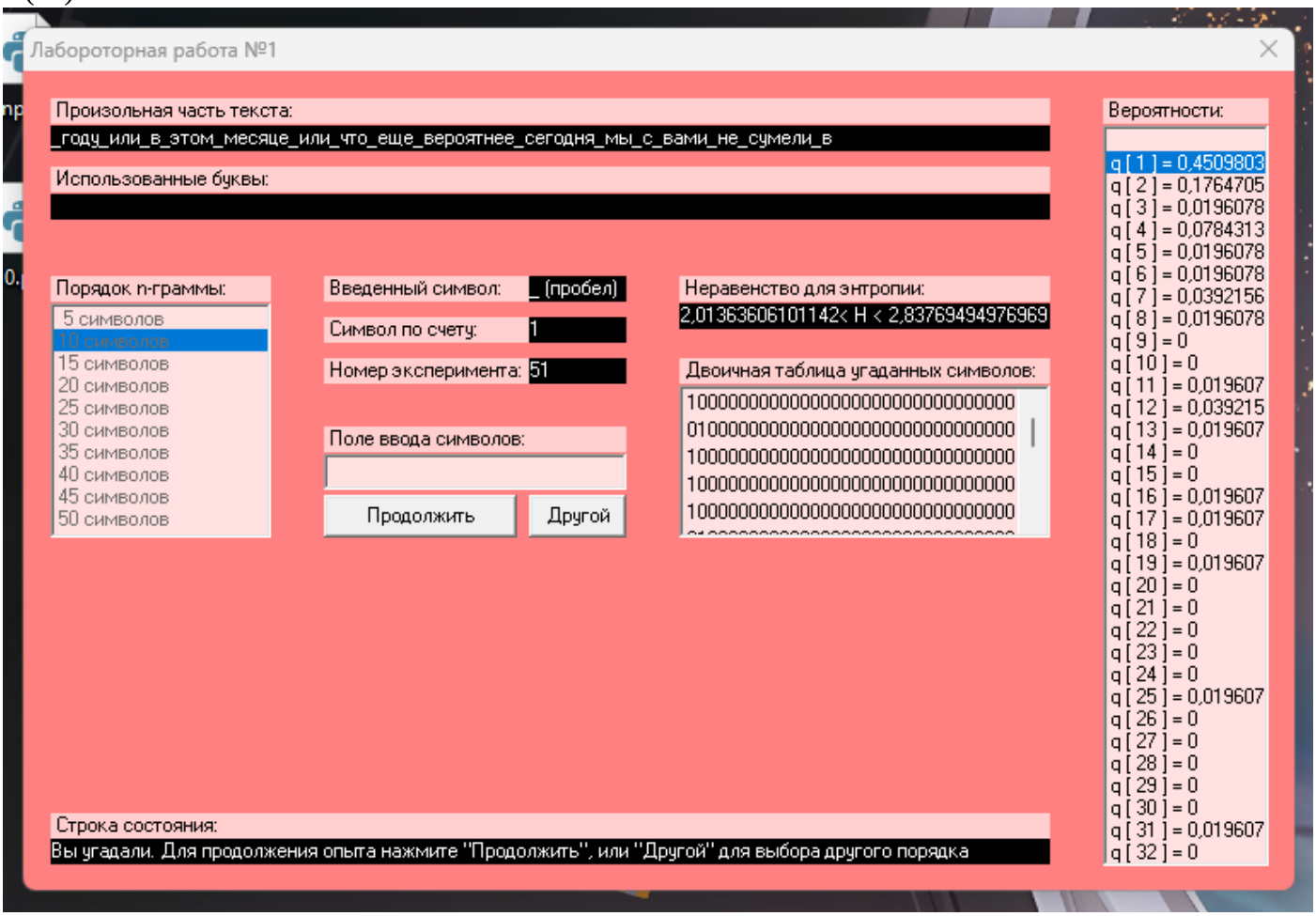
H2 block bigrams:	4.003559323053849
H2 block bigrams without spaces:	4.133327222431862
H2 block bigrams with redundancy:	0.20633494760258309
H2 block bigrams without spaces with redundancy:	0.17333455551362764

---H2 cross entropy---

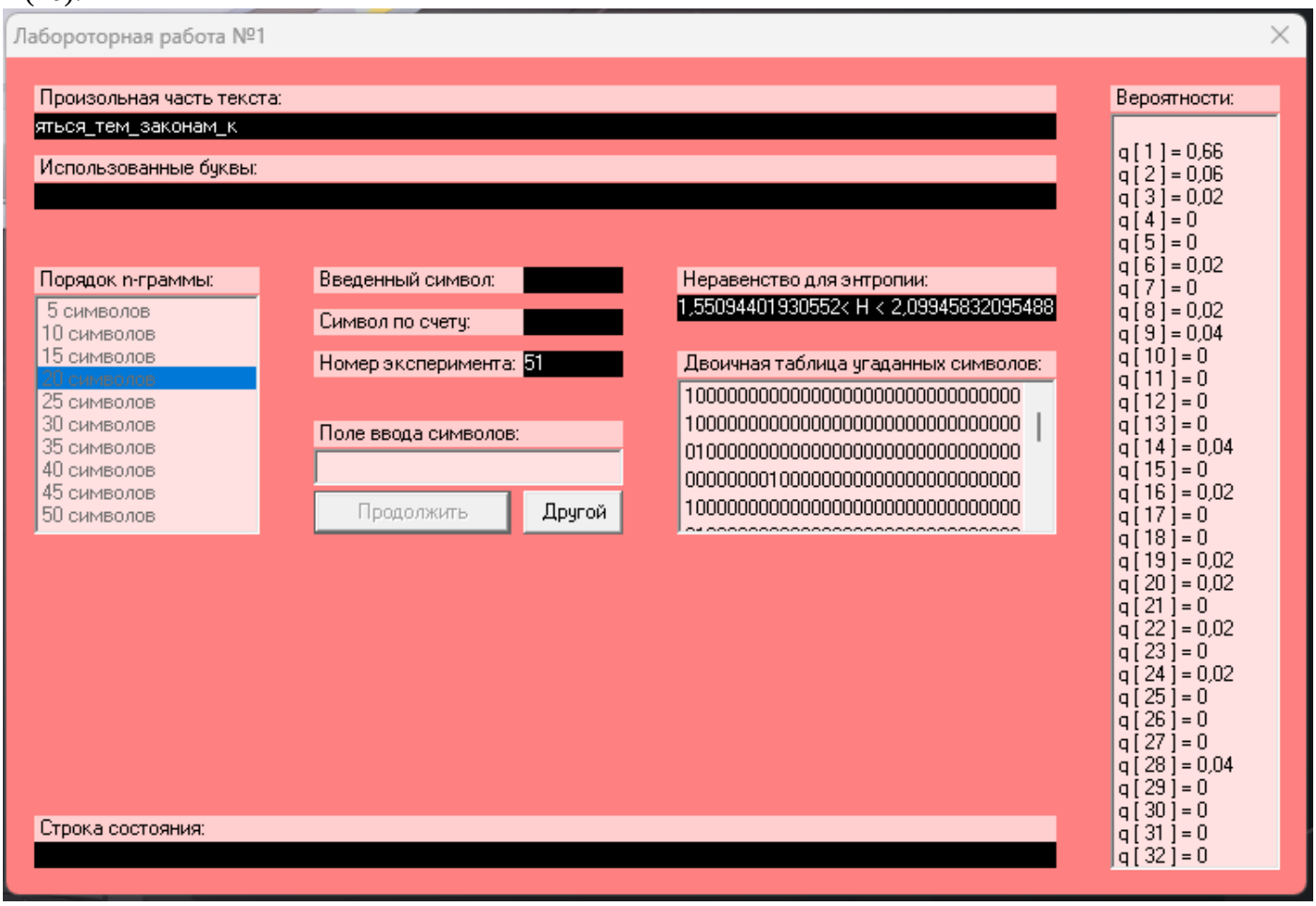
H2 cross bigrams:	4.0045129023504415
H2 cross bigrams without spaces:	4.133213304561436
H2 cross bigrams with redundancy:	0.2061459101733042
H2 cross bigrams without spaces with redundancy:	0.17335733908771278

CoolPinkProgram

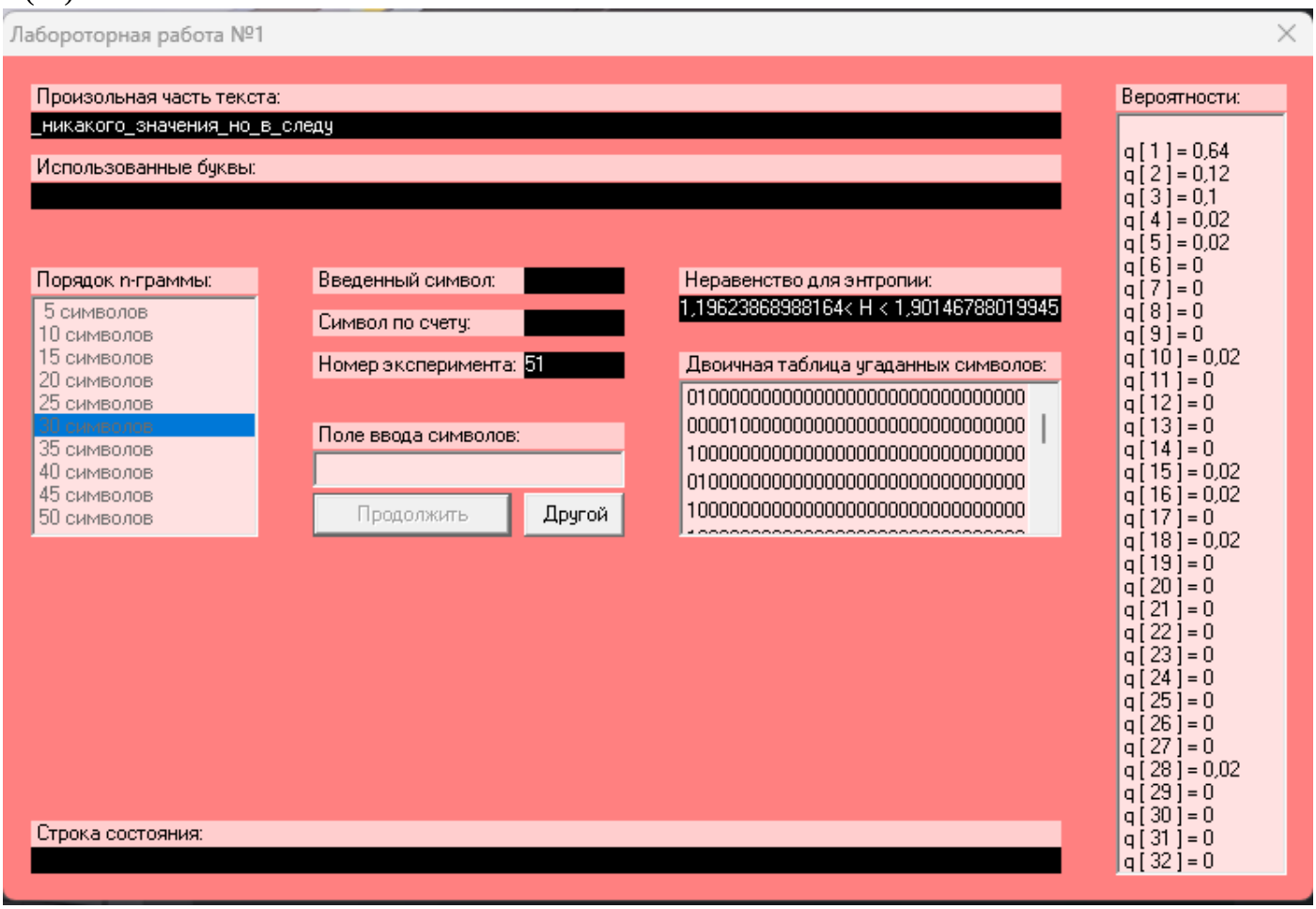
H(10):



H(20):


Лабороторная работа №1
Произольная часть текста:
яться_тем_законам_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:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,55094401930552< H < 2,09945832095488
Двоичная таблица угаданных символов:
1000000000000000000000000000000000
1000000000000000000000000000000000
0100000000000000000000000000000000
0000000010000000000000000000000000
1000000000000000000000000000000000
Строка состояния:
Вероятности:
q [1] = 0,66
q [2] = 0,06
q [3] = 0,02
q [4] = 0
q [5] = 0
q [6] = 0,02
q [7] = 0
q [8] = 0,02
q [9] = 0,04
q [10] = 0
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0,04
q [15] = 0
q [16] = 0,02
q [17] = 0
q [18] = 0
q [19] = 0,02
q [20] = 0,02
q [21] = 0
q [22] = 0,02
q [23] = 0
q [24] = 0,02
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0,04
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0

H(30):


Лабороторная работа №1
Произольная часть текста:
_никакого_значения_но_в_следу
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:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,19623868988164< H < 1,90146788019945
Двоичная таблица угаданных символов:
0100000000000000000000000000000000
0000100000000000000000000000000000
1000000000000000000000000000000000
0100000000000000000000000000000000
1000000000000000000000000000000000
Строка состояния:
Вероятности:
q [1] = 0,64
q [2] = 0,12
q [3] = 0,1
q [4] = 0,02
q [5] = 0,02
q [6] = 0
q [7] = 0
q [8] = 0
q [9] = 0
q [10] = 0,02
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0,02
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] = 0,02
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0

Результати:
2.0136 < H(10) < 2.8376
1.5509 < H(20) < 2.0994
1.1962 < H(30) < 1.9014

Оцінка надлишковості R російської мови у різних моделях відкритого тексту:

	H8	R
H(10)	2,4256	0,51488
H(20)	1,82515	0,63497
H(30)	1,5488	0,69024

Висновки:
Протягом даної лабораторної роботи нам необхідно було навчитися працювати з великими масивами інформації , використовуючи математичні функції та формули, а саме ентропія, надлишковість тексту, порівняння різних текстів відносно ентропії та зрозуміння явища ентропії