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

ъ	0.00023589313677991353	ъ	0.0002774369097796511
ф	0.0025313148139075336	ф	0.0029771114549431788
щ	0.0029468496471583045	щ	0.0034658272421704103
ц	0.002955922460111378	ц	0.003476497892546551
э	0.0049319811212908075	э	0.005800565544469936
ю	0.005387436331535102	ю	0.006336232193352185
ш	0.006958847535007449	ш	0.008184388838499706
х	0.007205628047331051	х	0.008474630528730726
ж	0.008564735427701475	ж	0.010073093955076561
й	0.00902019063794577	й	0.010608760603958812
ч	0.012536812938557096	ч	0.01474470468975084
з	0.013975761072914568	з	0.016437069839406713
г	0.014603599729267261	г	0.01717547884543563
б	0.014603599729267261	б	0.01717547884543563
я	0.01640546038174768	я	0.019294670010137117
ь	0.01673389621064894	ь	0.0196809475537534
ы	0.01733996011591426	ы	0.02039374699887958
у	0.02164228801826176	у	0.02545376940724537
п	0.022589489690562643	п	0.026567785306514432
д	0.023979444634973517	д	0.028202528944139146
м	0.026993433097984566	м	0.03174731899909299
к	0.0274416300578664	к	0.03227444912767433
в	0.03684832252761311	в	0.04333777943765672
р	0.039539318849494735	р	0.046502694339219976
л	0.04139743094228421	л	0.048688043536253535
с	0.04785545920228199	с	0.05628341247399029
н	0.053074141212889926	н	0.06242117057034626
т	0.0553024240741648	т	0.06504188230272635
и	0.059911413054326186	и	0.07046257269380568
е	0.0678392370127219	е	0.07978658699247719
а	0.07059555758786566	а	0.08302833057674865
о	0.09831118659691489	о	0.11562503334578242
	0.14974133410270787		

частоти біграм з пробілами

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

та	0.006419910796264911	ит	0.006090794236153794
д	0.006459831100820544	ет	0.00611213549136106
м	0.006477976693800376	ом	0.006197500512190125
й	0.006514267879760042	ло	0.006308475039267909
ос	0.0068481467905889696	ва	0.006381035306972615
ов	0.0069570203484679675	об	0.006385303558014068
ли	0.007015086246003433	од	0.006415181315304241
ен	0.0070259736017913325	ер	0.006419449566345694
пр	0.007080410380730832	ре	0.006517619340299119
ал	0.00715662187124613	те	0.007059687222563682
ор	0.0075739705097822895	ат	0.007209076009014546
т	0.00773365172800482	го	0.0072346855152632655
ла	0.00790422030201525	ан	0.0072944410298436116
ни	0.008205437145480478	ть	0.007328587038175237
ко	0.008299794228975609	ка	0.007473707573584648
к	0.008390522193874773	ас	0.007495048828791914
не	0.0087788378836432	та	0.007648705866284231
ро	0.008891340560118165	во	0.007721266133988936
о	0.009203444759371291	ак	0.007853581916273987
по	0.009217961233755158	ол	0.007951751690227412
ь	0.009486516009856685	ес	0.008203578501673154
я	0.009838540513665447	пр	0.008378576794372738
и	0.00986757346243318	ли	0.008583452844362493
на	0.009874831699625112	от	0.008660281363108653
ра	0.010281292982373371	ал	0.008664549614150106
но	0.010803886060192561	он	0.009121252475585604
ст	0.012132143466316336	ор	0.009304787270368094
н	0.013148296673186984	ла	0.009787099638052311
в	0.01392129893412787	ен	0.009855391654715565
то	0.014759625329796152	ни	0.009868196407839924
п	0.01489027359925095	не	0.009953561428668989
и	0.015877393857353865	ко	0.01015416922761729
е	0.017060486519638974	ос	0.01061087208905279
с	0.01716210184032604	ов	0.010751724373420746
а	0.017423398379235636	по	0.010760260875503653
о	0.021625917713364955	ро	0.010764529126545106

---H1 entropy---

H1: 4.401484367688899

H1 without spaces: 4.460162425027035

H1 with redundancy: 0.13483704843981703

H1 without spaces with redundancy: 0.12330319371318177

---H2 entropy---

H2 bigrams: 4.003559323053849

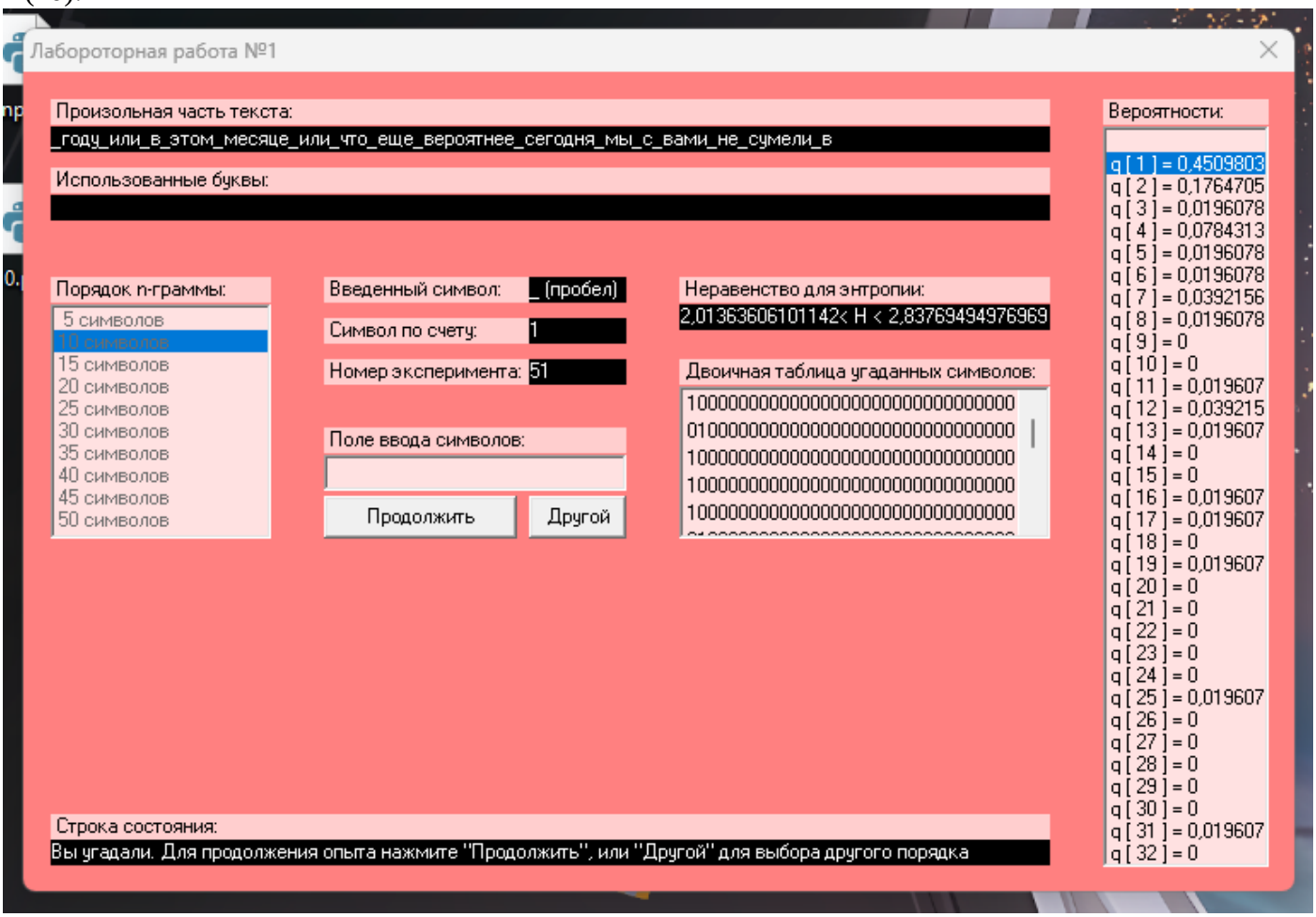
H2 bigrams without spaces: 4.133327222431862

H2 bigrams with redundancy: 0.2130538447196012

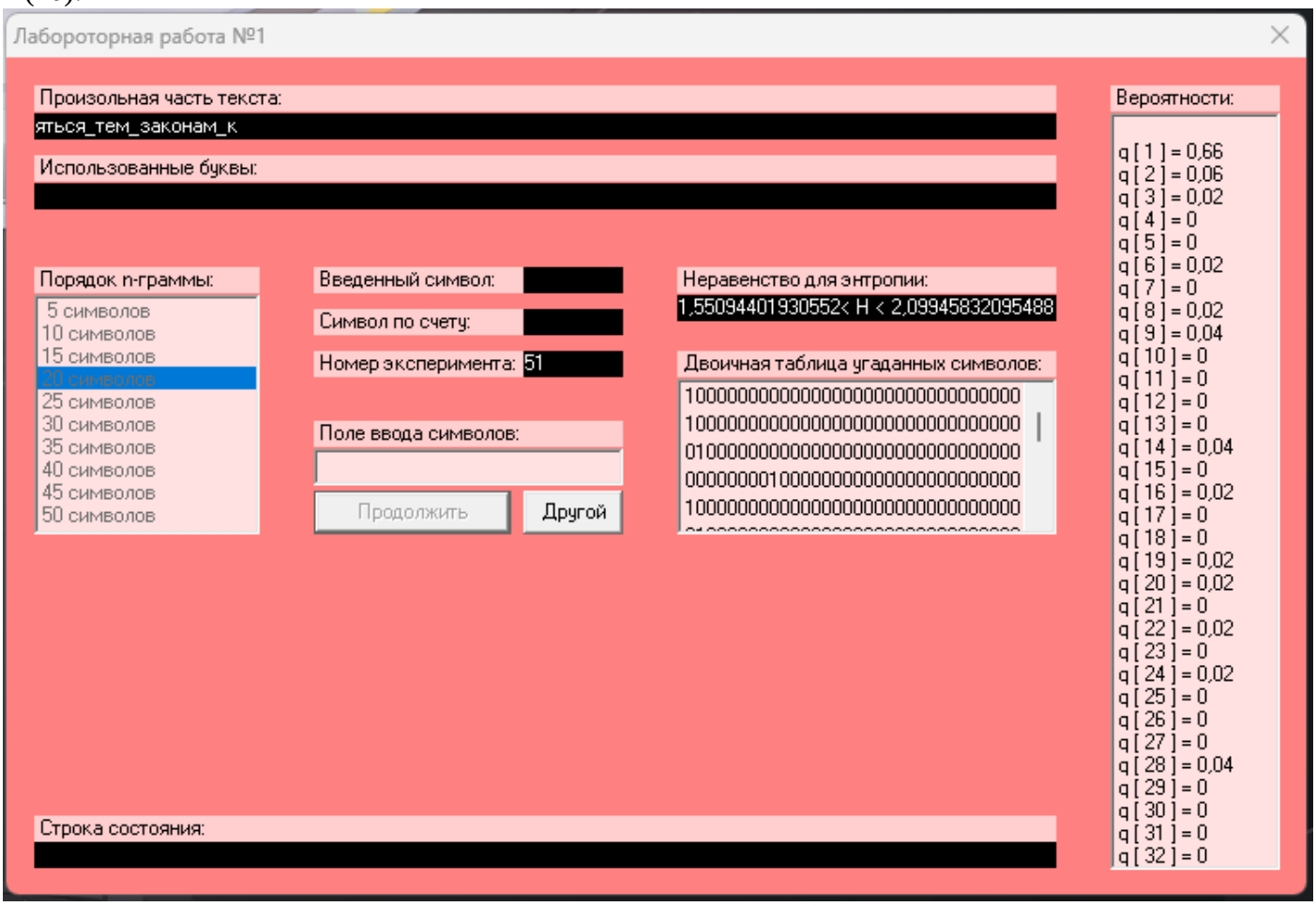
H2 bigrams without spaces with redundancy: 0.1806097769859547

CoolPinkProgram

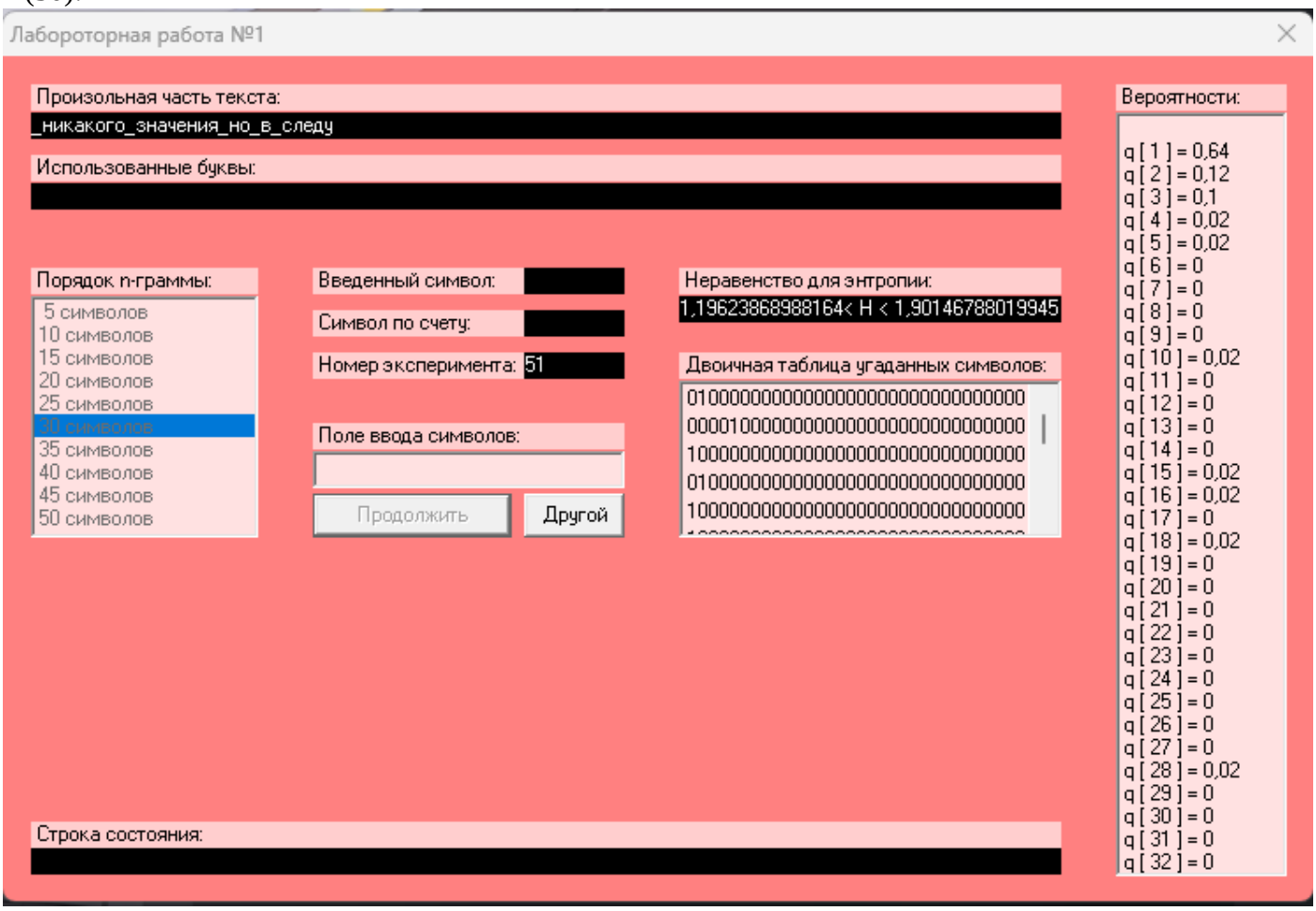
H(10):



H(20):



H(30):



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