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

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

з дисципліни «Криптографія»

Виконав:

студент 3 курсу

НН ФТІ групи ФБ-25

Черняк Денис

Тема: «Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Мета: 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.

Хід роботи

Завдання 1

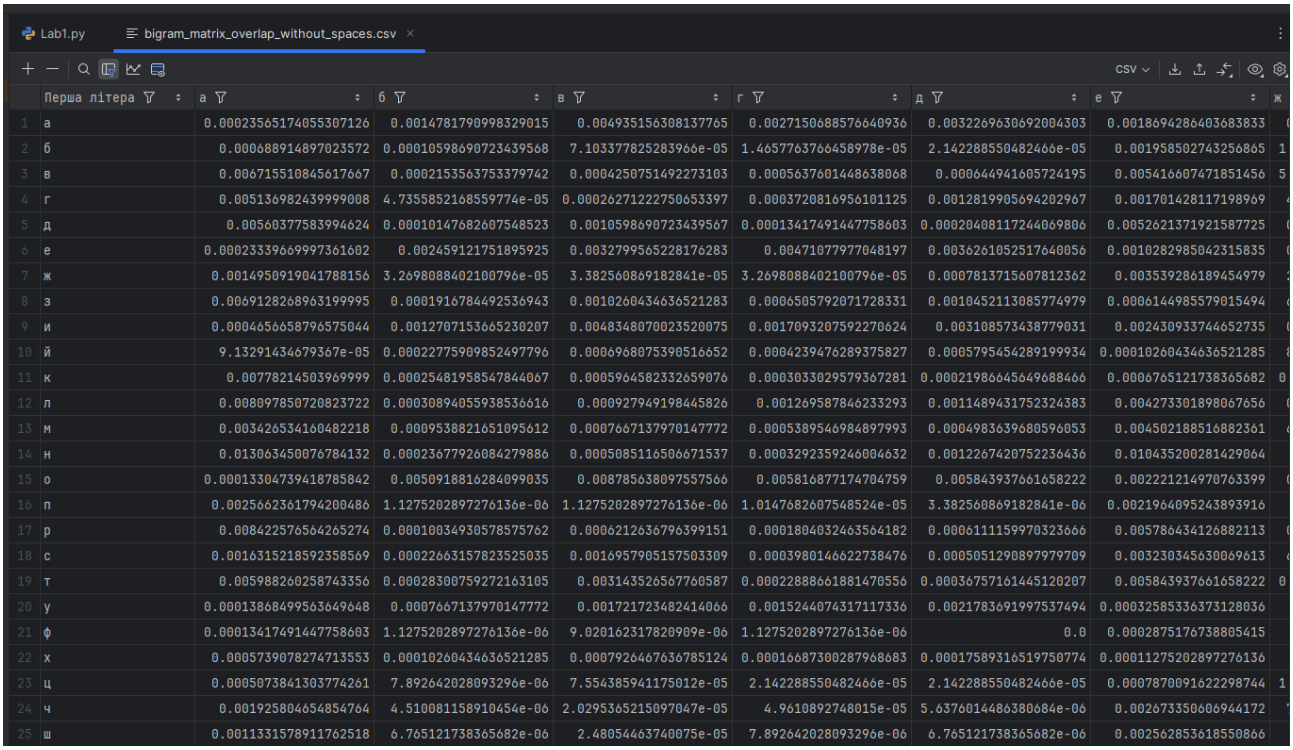
Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними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

Виконання

Частоти біграм з перетином:

Lab1.py | bigram_matrix_overlap_with_spaces.csv

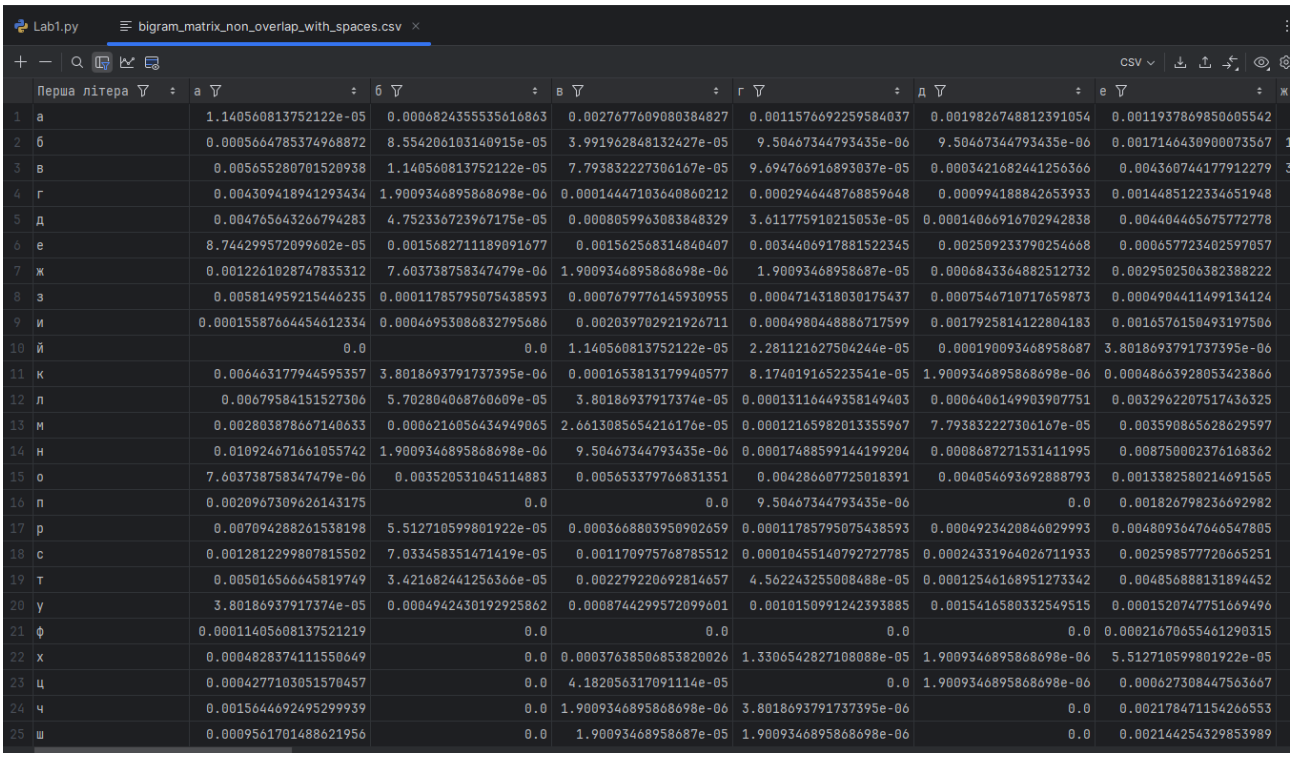
	Перша літера	а	б	в	г	д	е
1	а	1.2356075482314655e-05	0.0006985934984231747	0.002823838481381295	0.0011633720300271643	0.0019959814240662135	0.0011291552056146006
2	б	0.0005731318089104413	8.554206103140915e-05	4.6572899894878315e-05	9.50467344793435e-06	1.2356075482314655e-05	0.0016490608432166095
3	в	0.0056106087363156 46	1.0455140792727785e-05	5.98794427219864e-05	0.00011025421199603845	0.0003269076660894164	0.004433930163461374
4	г	0.004319874082086162	1.9009346895868698e-06	0.0001482729057877 7586	0.0002879916054724108	0.0010322075364456703	0.0014038402682599034
5	д	0.00470671429141709	4.6572899894878315e-05	0.0007945907002473116	2.946448768859648e-05	0.0001207093527887 6623	0.004403515208427984
6	е	7.223551820430106e-05	0.0015986860739425576	0.001549261772013299	0.0034739581452200045	0.0024769179005316912	0.0006482187291491226
7	ж	0.0012508150257481603	1.1405608137521 22e-05	2.851402034380305e-06	2.091028158545557e-05	0.0006434663924251554	0.0029797151259274186
8	з	0.0058111573460670 61	0.0001207093527887 6623	0.0007299589208013581	0.0004818869438102715 3	0.000817401916522354	0.0004799860091206846 3
9	и	0.0001786878608211 6578	0.0004790355417758912	0.0021223935809237403	0.0005085000294644877	0.0017830767388324839	0.0016310019636655343
10	й	9.504673447934349e-07	9.504673447934349e-07	1.0455140792727785e-05	3.421682441256366e-05	0.0001929448709930673	4.752336723967175e-06
11	к	0.0065259087893517 24	1.9009346895868698e-06	0.0001748859914419 9204	7.413645289388793e-05	9.504673447934349e-07	0.0004847383458446518
12	л	0.0067359620725510735	5.417663865322579e-05	3.32663570677702 25e-05	0.0001254616895127 3342	0.0006225561108396999	0.0032173319621257774
13	м	0.0028105319385541873	0.0006102000353573852	3.0414955033389916e-05	0.0001178579507543 8593	7.508692023868136e-05	0.0036849618957641474
14	н	0.010974095962985	2.851402034380305e-06	8.554206103140915e-06	0.0001596785139252 9708	0.000869677620485993	0.008722438823169353
15	о	5.702804068760 61e-06	0.0035347880552867843	0.005626766681177135	0.004218174076193264	0.00411742453764516	0.001411444007018251
16	п	0.0021632636767498 58	0.0	0.0	6.653271413554044e-06	0.0	0.0018515103876576113
17	р	0.0070790807840215 04	4.562243255008488e-05	0.0004077504909163 8357	0.0001150065487200056 2	0.0004533729234664 6845	0.004853086262515279
18	с	0.0013287533480212 22	7.413645289388793e-05	0.0011700253014407184	9.124486510016976e-05	0.0002404682382327 3904	0.0026346954797674014
19	т	0.0049871021581311 53	3.611775910215053e-05	0.0022041337725759755	4.277103051570457 5e-05	0.0001140560813752 1219	0.004791305885103706
20	у	2.661308565421617 6e-05	0.0004723822703623 3715	0.0008972411734850025	0.000958071083551782 4	0.0015464103699789186	0.0001843906648899264
21	ф	0.0001112046793408 3188	0.0	0.0	0.0	0.0	0.0002423691729223259
22	х	0.0004514719887768816	1.9009346895868698e-06	0.0003820878726069 6084	1.330654282710808 8e-05	4.752336723967175e-06	5.607757334281266e-05
23	ц	0.0004229579684330 7853	0.0	4.847383458446518 5e-05	1.9009346895868698e-06	2.851402034380305e-06	0.0006624757393210241
24	ч	0.0016186458881832196	0.0	9.504673447934349e-07	2.851402034380305e-06	0.0	0.0022516571398156476
25	ш	0.0009542692141726087	0.0	1.8058879551075265e-05	9.504673447934349e-07	0.0	0.002158511340025891



Lab1.py | bigram_matrix_overlap_without_spaces.csv

	Перша літера	а	б	в	г	д	е	ж
1	а	0.00023565174055307126	0.0014781790998329015	0.004935156308137765	0.0027150688576640936	0.0032269630692004303	0.0018694286403683833	0
2	б	0.000688914897023572	0.0001059869072343956 8	7.103377825283966e-05	1.4657763766458978e-05	2.142288550482466e-05	0.00195850274325 6865	1
3	в	0.006715510845617667	0.0002153563753379742	0.0004250751492273103	0.0005637601448638068	0.000644941605724195	0.005416607471851456	5
4	г	0.005136982439999008	4.7355852168559774e-05	0.00026271222750653397	0.0003720816956101125	0.0012819905694202967	0.001701428117198969	4
5	д	0.00560377583994624	0.0001014768260754852 3	0.0010598690723439567	0.0001341749144775860 3	0.0002040811724406980 6	0.0052621371921587725	0
6	е	0.00023339669997361602	0.002459121751895925	0.0032799565228176283	0.00471079779770481 97	0.0036261052517640056	0.0010282985042315835	0
7	ж	0.0014950919041788156	3.2698088402100796e-05	3.382560869182841e-05	3.2698088402100796e-05	0.0007813715607812362	0.0035392861894549 79	2
8	з	0.0069128268963199995	0.0001916784492536943	0.0010260043463652128 3	0.0006505792071728331	0.0010452113085774979	0.0006144985579015494	0
9	и	0.0004656658796575044	0.0012707153665230207	0.0048348070023520075	0.0017093207592270624	0.003108573438779031	0.002430933744652735	0
10	й	9.132914346793 67e-05	0.0002277590985 2497796	0.0006968075390516652	0.0004239476289375827	0.0005795454289199934	0.0001026043463652128 5	8
11	к	0.00778214503969999	0.0002548195854784406 7	0.0005960458332659076	0.0003033029579367281	0.0002198664564968846 6	0.0006765121738365682	0
12	л	0.008097850720823722	0.0003089405593853661 6	0.000927949198445826	0.001269587846233293	0.0011489431752324383	0.004273301898067656	0
13	м	0.003426534160482218	0.0009538821651095612	0.0007667137970147772	0.0005389546984897993	0.0004983639680596053	0.004502188516882361	4
14	н	0.013063450076784132	0.0002367792608427988 6	0.0005085116506671537	0.0003292359246004632	0.0012267420752236436	0.010435200281429064	
15	о	0.00013304739418785842	0.0050918816284099035	0.008785638097557566	0.005816877174704759	0.005843937661658222	0.002221214970763399	0
16	п	0.0025662361794200486	1.1275202897276136e-06	1.1275202897276136e-06	1.0147682607548524e-05	3.382560869182841e-06	0.0021964095243893916	
17	р	0.008422576564265274	0.0001003493057857576 2	0.0006212636796399151	0.0001804032463564182	0.0006111159970323666	0.005786434126882113	0
18	с	0.0016315218592358569	0.0002266315782352503 5	0.0016957905157503309	0.0003980146622738476	0.0005051290897979709	0.003230345630069613	4
19	т	0.005988260258743356	0.0002830075927216310 5	0.003143526567760587	0.0002288861881470556	0.0003675716144512020 7	0.005843937661658222	0
20	у	0.0001386849956364964 8	0.0007667137970147772	0.0017217234824140 66	0.0015244074317117336	0.0021783691997537494	0.0003258533637312803 6	
21	ф	0.0001341749144775860 3	1.1275202897276136e-06	9.020162317820989e-06	1.1275202897276136e-06	0.0	0.0002875176738805415	
22	х	0.0005739078274713553	0.0001026043463652128 5	0.0007926467636785124	0.0001668730028796868 3	0.0001758931651975077 4	0.0001127520289727613 6	
23	ц	0.0005073841303774261	7.892642028093296e-06	7.554385941175012e-05	2.142288550482466e-05	2.142288550482466e-05	0.0007870091622298744	1
24	ч	0.001925804654854764	4.5100811581 9104540e-06	2.0295365215097047e-05	4.9610892748015e-05	5.6376014486380684e-06	0.002673350606944172	1
25	ш	0.0011331578911762518	6.765121738365682e-06	2.480544637400 75e-05	7.892642028093296e-06	6.765121738365682e-06	0.002562853618550866	

Частоти біграм без перетину:



Lab1.py | bigram_matrix_non_overlap_with_spaces.csv

	Перша літера	а	б	в	г	д	е	ж
1	а	1.1405608137521 22e-05	0.0006824355535616863	0.0027677609080384827	0.0011576692259584037	0.0019826748812391054	0.0011937869850605542	
2	б	0.0005664785374968872	8.554206103140915e-05	3.991962848132427e-05	9.504673447934 35e-06	9.504673447934 35e-06	0.0017146430900073567	1
3	в	0.005655280701520938	1.1405608137521 22e-05	7.7938322273061 67e-05	9.6947669168930 37e-05	0.0003421682441256366	0.004360744177912279	3
4	г	0.004309418941293434	1.9009346895868698e-06	0.0001444710364086021 2	0.0002946448768859648	0.000994188842653933	0.0014485122334651948	
5	д	0.004765643266794283	4.752336723967175e-05	0.0008059963083848329	3.611775910215053e-05	0.0001406691670294 2838	0.004404465675772778	
6	е	8.7442995720996 02e-05	0.0015682711189091677	0.001562568314840407	0.0034406917881522345	0.002509233790254668	0.000657723402597057	
7	ж	0.0012261028747835312	7.6037387583474 79e-06	1.9009346895868698e-06	1.9009346895868 7e-05	0.0006843366488251 2732	0.0029502506382388222	
8	з	0.005814595215446235	0.0001178579507543859 3	0.0007679776145930955	0.0004714318030175437	0.0007546710717659873	0.0004904411499134124	
9	и	0.0001558766445461233 4	0.0004695308683279568 6	0.002039702921926711	0.0004980448886717599	0.0017925814122804183	0.0016576150493197506	
10	й	0.0	0.0	1.140560813752122e-05	2.281121627504244e-05	0.0001909034689589 87	3.8018693791737395e-06	
11	к	0.0064631779445955 57	3.8018693791737395e-06	0.0001653813179940577	8.174019165223541e-05	1.9009346895868698e-06	0.0004866392805342 3866	
12	л	0.006795841515273 06	5.7028040687606 09e-05	3.801869379173 74e-05	0.0001311644935814 9403	0.0006406149903907751	0.0032962207517436325	
13	м	0.002803878667140633	0.0006216056434949065	2.661308565421617 6e-05	0.0001216598201335 5967	7.7938322273061 67e-05	0.0035908656286295 97	
14	н	0.010924671661055742	1.9009346895868698e-06	9.504673447934 35e-06	0.0001748859914419 9204	0.0008687271531411995	0.0087500023761683 62	
15	о	7.6037387583474 79e-06	0.0033520531045114883	0.005653379766831351	0.004286607725018391	0.004054693692888793	0.0013382580214691565	
16	п	0.0020967309626143175	0.0	0.0	9.504673447934 35e-06	0.0	0.001826798236692982	
17	р	0.007094288261538198	5.5127105998019 22e-05	0.0003668803950902659	0.0001178579507543859 3	0.0004923420846029993	0.0048093647646547805	
18	с	0.0012812299807815502	7.0334583514714 19e-05	0.001170975768785512	0.0001045514079272778 5	0.0002433196402671193 3	0.002598577720665251	
19	т	0.0050165666458197 49	3.4216824412563 66e-05	0.002279220692814657	4.562243255008488e-05	0.0001254616895127334 2	0.004856888131894452	
20	у	3.801869379173 74e-05	0.0004942430192925862	0.0008744299572099601	0.0010150991242393885	0.0015416580332549515	0.0001520747751669496	
21	ф	0.0001140560813752 1219	0.0	0.0	0.0	0.0	0.0002167065546129031 5	
22	х	0.0004828374111550649	0.0	0.0003763850685382002 6	1.3306542827108088e-05	1.9009346895868698e-06	5.5127105998019 22e-05	
23	ц	0.0004277103051570457	0.0	4.182056317091114e-05	0.0	1.9009346895868698e-06	0.000627308447563667	
24	ч	0.0015644692495299939	0.0	1.9009346895868698e-06	3.8018693791737395e-06	0.0	0.002178471154266553	
25	ш	0.0009561701488621956	0.0	1.900934689586 87e-05	1.9009346895868698e-06	0.0	0.0021442543298539 89	

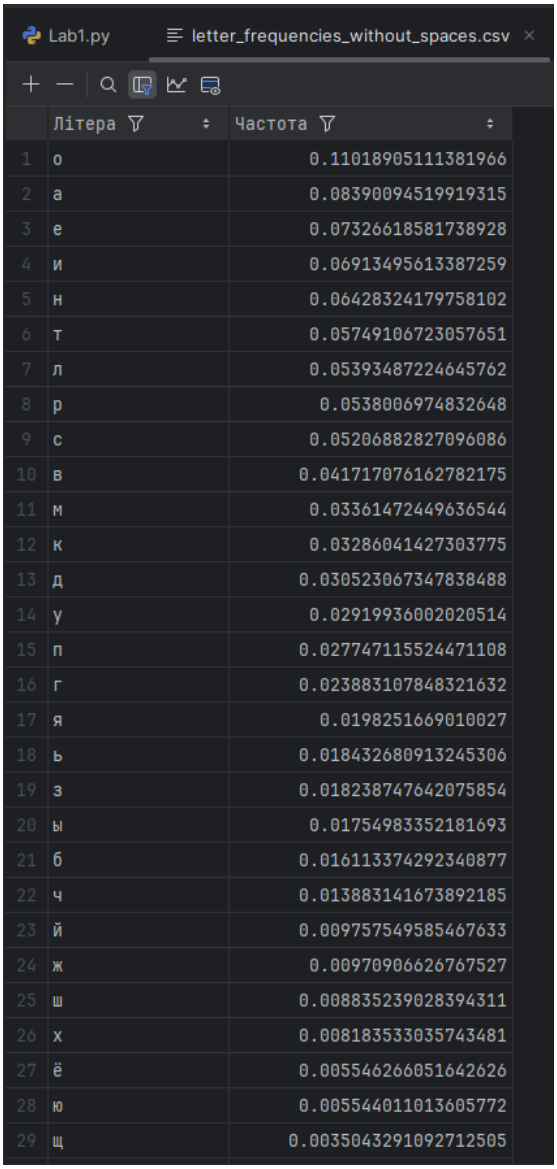
Lab1.py | bigram_matrix_non_overlap_without_spaces.csv

	Перша літера	а	б	в	г	д	е	ж
1	а	0.0002300141391044332	0.001468031417225353	0.004927263666109672	0.002667713005495534	0.0032179429068826095	0.0018130526258820028	
2	б	0.000703572660790031	0.00012177219129058228	6.990625796311204e-05	1.578528405618659e-05	2.48054463740075e-05	0.0019731605070233238	1
3	в	0.006920719538348092	0.0001894234086742391	0.0004329677912554036	0.0005637601448638068	0.0006584718492009264	0.005518084297926941	
4	г	0.005175318129849746	6.314113622474637e-05	0.00023677926084279886	0.0003517863303950155	0.0013079235360840319	0.0016935354751708757	
5	д	0.005587990555890053	9.245666375766432e-05	0.001062124112923412	0.00014432259708513454	0.0001646179623002316	0.005233949184915583	
6	е	0.0002503090431953025	0.0024647593533445636	0.003364520544547199	0.004776175947286171	0.0036576758198763786	0.001071144275241233	
7	ж	0.0015153872693939128	3.157056811237318e-05	3.382560869182841e-05	3.833568985073886e-05	0.000773478918753143	0.003549433872062528	1
8	з	0.006848558239805525 5	0.0001961885304126048	0.001012513220175397	0.0006674920115187473	0.0010260434636521283	0.0005727803071816277	
9	и	0.0004713034811061425	0.001231252156382554	0.0047378402574354325	0.001666474988217412 9	0.003073620309797475	0.002521135367830944	
10	й	8.794658259875387e-05	0.0002300141391044332	0.0006697470520982025	0.0004036522637224857	0.0005524849419665307	0.00011275202897276136	0
11	к	0.00784303113534528	0.0002300141391044332	0.0005569950231254411	0.00029766535648809	0.00021197381446879136	0.0007080827419489414	0
12	л	0.008113636004879908	0.0003044304782264557	0.000983197692642479	0.0012876281708689347	0.00109143964045633	0.004239476289375827	
13	м	0.003393836072080117	0.0009110363940999118	0.0007915192433887848	0.0005434647796487098	0.0004938538869006947	0.004496550915433723	
14	н	0.013279933972411834	0.0002232490173660675	0.0005412097390692545	0.0002954103159086348	0.0012041916694290914	0.010298770326372024	
15	о	0.0001307923536084031 8	0.00515953284579356	0.008742792326547916	0.0057120177877600905	0.005793199248620479	0.0022370002548195854	
16	п	0.002640652518542071	0.0	0.0	9.020162317820909e-06	2.255040579455227e-06	0.0022257250519223093	
17	р	0.008478952578751654	9.471170433711955e-05	0.000644941605724195	0.0001871683680947838 6	0.0006133710376118218	0.005763883721087561	
18	с	0.0016551997853201368	0.0002300141391044332	0.0016957905157503309	0.0003811018579279334	0.0004938538869006947	0.003242748353256617	
19	т	0.005894676074695964	0.0002886451941702691	0.0030871505532742063	0.0002074637333098809	0.0003901220202457543	0.005842810141368494	
20	у	0.0001240272318700375	0.0007216129854256728	0.0017611866925545325	0.00158754856793648	0.002200919605548302	0.0003157056811237318	
21	ф	0.0001307923536084031 8	2.255040579455227e-06	6.765121738365682e-06	0.0	0.0	0.0002954103159086348	
22	х	0.0005682702260227173	8.794658259875387e-05	0.0007802440404915086	0.0001894234086742391	0.00019844357099206	0.0001240272318700375	
23	ц	0.0005276794955925232	1.3530243476731363e-05	7.44163391220225e-05	2.0295365215097047e-05	1.8040324635641818e-05	0.0007464184317996802	
24	ч	0.0018987441679013013	2.255040579455227e-06	2.48054463740075e-05	4.7355852168559774e-05	4.5100811589104 54e-06	0.002699283573607907	
25	ш	0.0011523257361016211	9.020162317820909e-06	1.3530243476731363e-05	2.255040579455227e-06	9.020162317820909e-06	0.002559471057681683	

Частоти букв:

Lab1.py | letter_frequencies_with_spaces.csv

	Літера	Частота
1		0.1570284617175879
2	о	0.09288623391929589
3	а	0.07072610883791221
4	е	0.06176130936256968
5	и	0.05827880032125766
6	н	0.05418894322388712
7	т	0.048463333380856656
8	л	0.045465562224661756
9	р	0.045352456718134426
10	с	0.043892540264134625
11	в	0.035166307865585035
12	м	0.028336256017640657
13	к	0.027700393968339964
14	д	0.025730077035305075
15	у	0.02461422943309429
16	п	0.023390028656563207
17	г	0.020132780161864434
18	я	0.01671205143924381
19	ь	0.01553822538410725
20	з	0.015374745156185398
21	ы	0.01479401016048626
22	б	0.013583115914134861
23	ч	0.011703093293033557
24	й	0.008225336583928563
25	ж	0.0081844665269481
26	ш	0.007447855034858356
27	х	0.006898485431725619



Lab1.py letter_frequencies_without_spaces.csv

	Літера	Частота
1	о	0.11018905111381966
2	а	0.08390094519919315
3	е	0.07326618581738928
4	и	0.06913495613387259
5	н	0.06428324179758102
6	т	0.05749106723057651
7	л	0.05393487224645762
8	р	0.0538006974832648
9	с	0.05206882827096086
10	в	0.041717076162782175
11	м	0.03361472449636544
12	к	0.03286041427303775
13	д	0.030523067347838488
14	у	0.02919936002020514
15	п	0.027747115524471108
16	г	0.023883107848321632
17	я	0.0198251669010027
18	ь	0.018432680913245306
19	з	0.018238747642075854
20	ы	0.01754983352181693
21	б	0.016113374292340877
22	ч	0.013883141673892185
23	й	0.009757549585467633
24	ж	0.00970906626767527
25	ш	0.008835239028394311
26	х	0.008183533035743481
27	ё	0.005546266051642626
28	ю	0.005544011013605772
29	щ	0.0035043291092712505

Ентропія та надлишковість:

З пробілами:

Тексти	H	R
Текст з пробілами монограми	4.406696524751244	0.12641708389912865
Текст з пробілами біграми(з перетином)	2.9401289580442764	0.70857461882675
Текст з пробілами біграми(без перетину)	2.939925017413529	0.7085948334081171

Без пробілів:

Тексти	H	R
Текст без пробілів монограми	4.48359558232663	0.1111726252474392
Текст без пробілів біграми(з перетином)	4.178746528108722	0.585802930021636
Текст без пробілів біграми(без перетину)	4.178337155941605	0.5858435069628336

Результат у програмі:

```
Ентропія та надлишковість для тексту з пробілами:
H1 (ентропія букв): 4.406696524751244, R: 0.12641708389912865
H2 (ентропія біграм з перетином): 2.9401289580442764, R: 0.70857461882675
H2 (ентропія біграм без перетину): 2.939925017413529, R: 0.7085948334081171

Ентропія та надлишковість для тексту без пробілів:
H1 (ентропія букв): 4.48359558232663, R: 0.1111726252474392
H2 (ентропія біграм з перетином): 4.178746528108722, R: 0.585802930021636
H2 (ентропія біграм без перетину): 4.178337155941605, R: 0.5858435069628336
```

Завдання 2

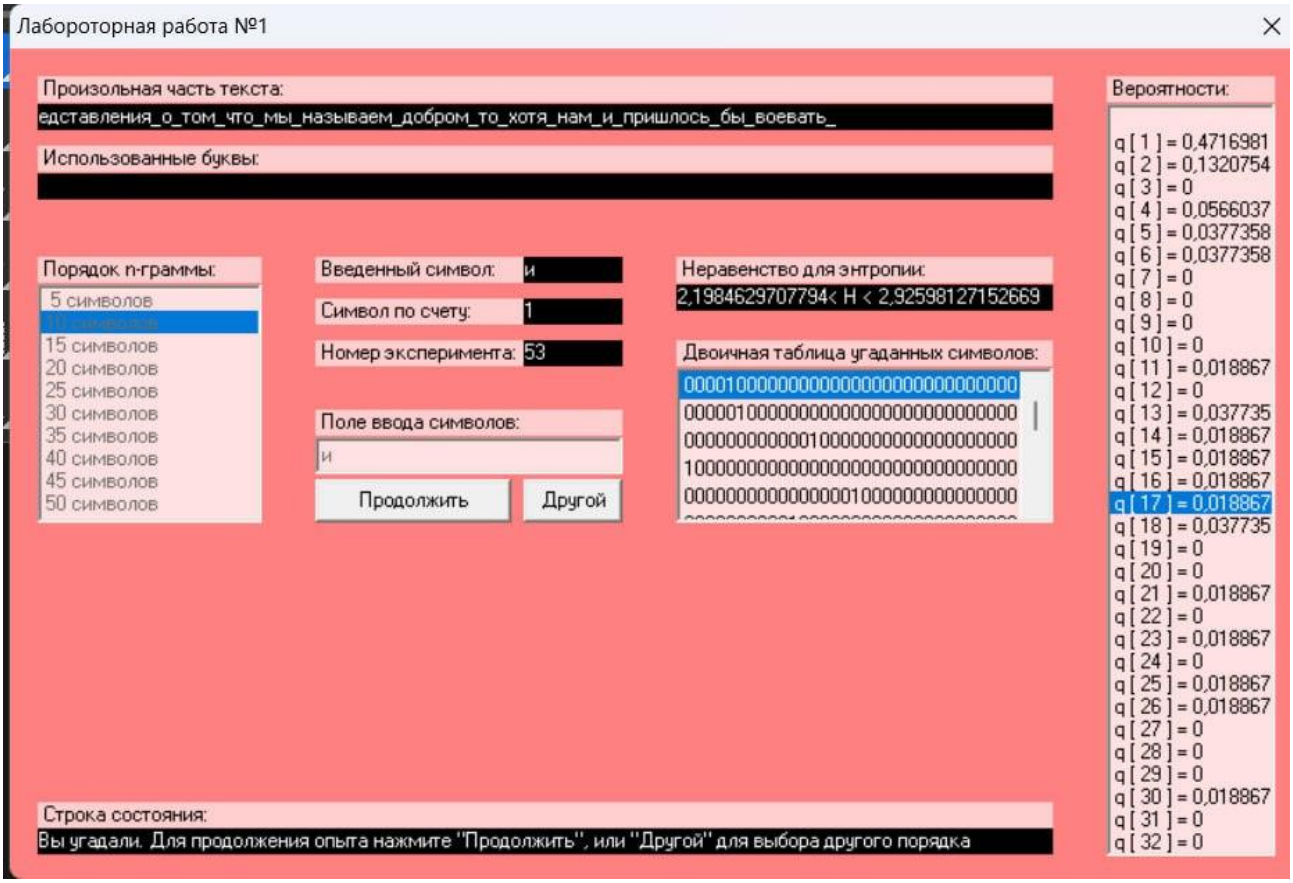
За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}, H^{(20)}, H^{(30)}$.

Під час виконання були проблеми з ієроглєфами, і на жаль, повністю прибрати її не вдалось, але залишок символів не заважав виконання роботи, оскільки одна кнопка відповідала за продовження дії, а інша до перехід до наступного значення

Виконання

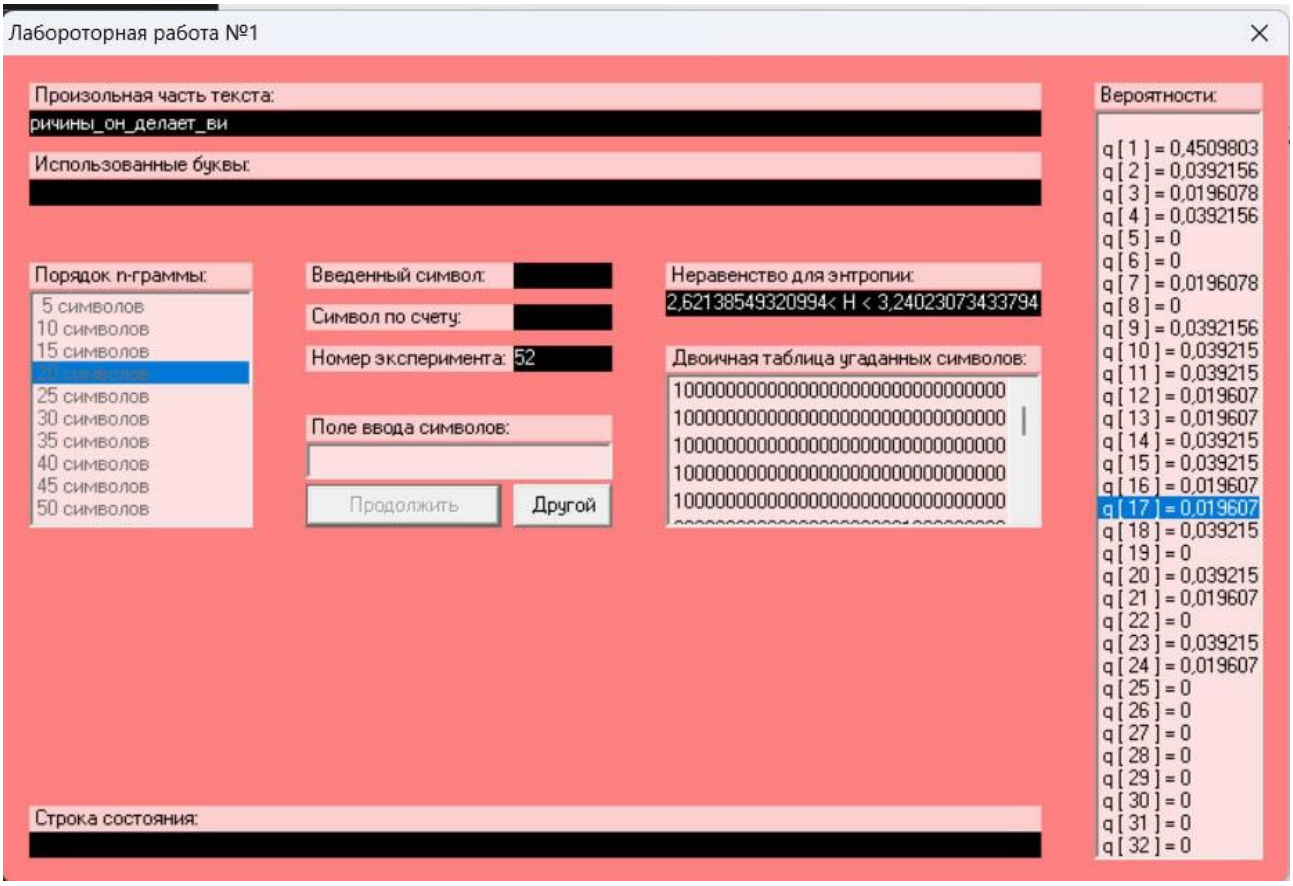
$H^{(10)}$

2,1984629707794 < H < 2,92598127152669



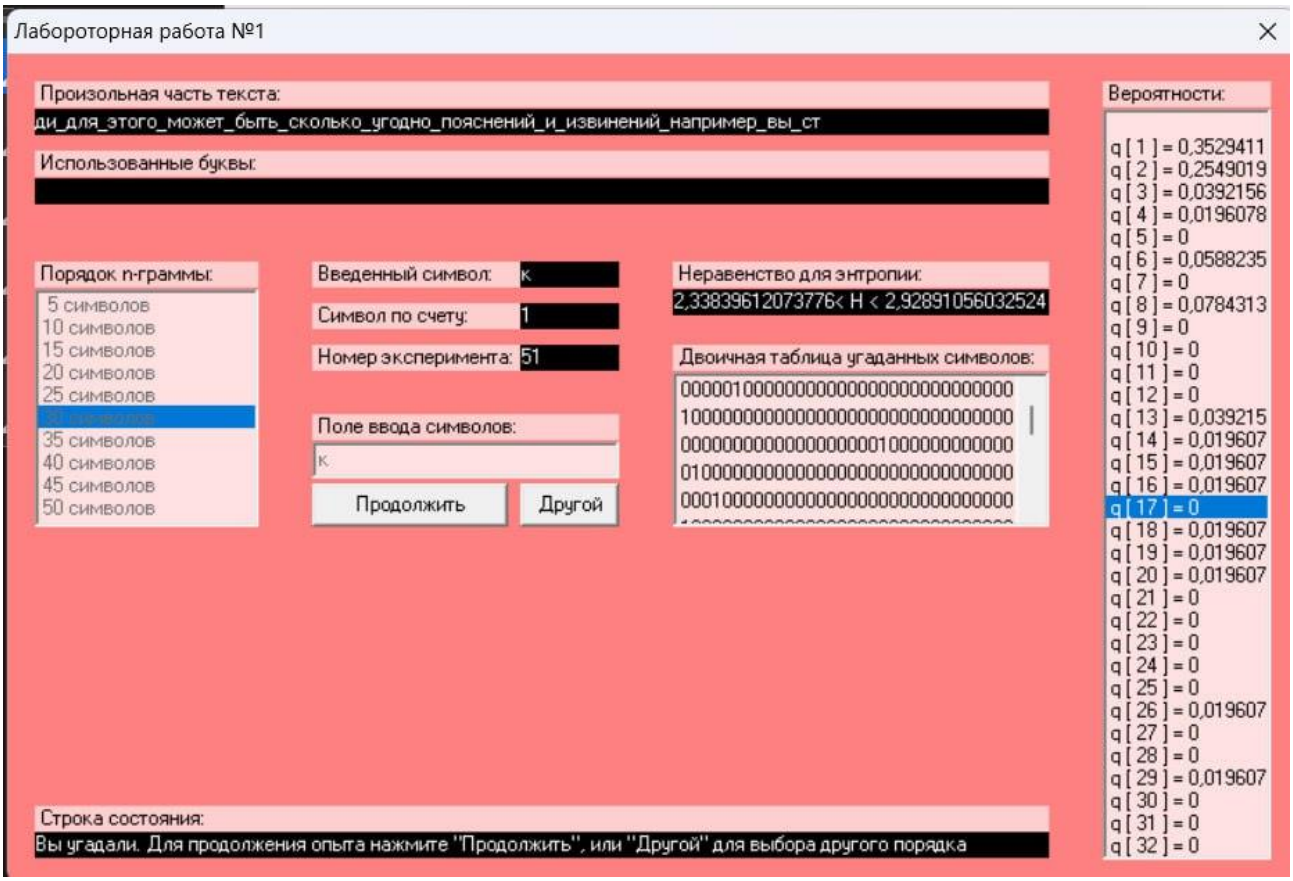
$H^{(20)}$

2,62138549320994 < H < 3,24023073433794



$H^{(30)}$

2,33839612073776 < H < 2,92891056032524



Завдання 3

Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Виконання

Для обрахунків використаємо цю формулу

$$R = 1 - \frac{H_\infty}{H_0}$$

$H_{(0)}$ можна знайти за наступною формулою, 32 це кількість літер у нашому алфавіті, це вказано у методичці, що у наданому тексті використовується лише 32 символи

$H_{(0)} = \log_2 32 = 5$

$H^{(10)}$

$$R = 1 - \frac{2{,}1984629707794}{5} \approx 0{,}56030740584412$$

$$R = 1 - \frac{2{,}92598127152669}{5} \approx 0{,}414803745694662$$

$H^{(20)}$

$$R = 1 - \frac{2{,}62138549320994}{5} \approx 0{,}475722901358012$$

$$R = 1 - \frac{3{,}24023073433794}{5} \approx 0{,}351953853132412$$

$H^{(30)}$

$$R = 1 - \frac{2{,}33839612073776}{5} \approx 0{,}532320775852448$$

$$R = 1 - \frac{2{,}92891056032524}{5} \approx 0{,}414217887934952$$

Висновки:

У ході виконання роботи я дослідив та оцінив ентропію для символів джерела відкритого тексту російською мовою. Було розроблено програму для підрахунку частот символів та біграм у тексті, а також обчислено ентропії H1 та H2 (з перетином та без) як з урахуванням пробілів, так і без них. Це дало можливість більш детально вивчити вплив пробілів на загальну ентропію тексту. Зокрема, результати показали, що включення пробілів впливає на розподіл частот окремих символів та біграм, проте не призводить до суттєвих змін значень H1 та H2 , що свідчить про низький внесок пробілів у загальну ентропію при обробці тексту значного обсягу.

Додатково, за допомогою програми CoolPinkProgram було обчислено ентропії H10, H20, H30 для тексту, що включає лише літери та пробіли, без урахування розділових знаків. Це дозволило оцінити надлишковість російської мови, яка, за нашими підрахунками, варіюється від 47% до 65% залежно від значення ентропії HHH. Високий рівень надлишковості свідчить про значну передбачуваність структури російської мови, що можна враховувати при проєктуванні систем стиснення тексту або криптографічного аналізу.

Результати також вказали на помітну різницю у значеннях H між текстом із пробілами та без них, а також між різними завданнями, що підтверджує залежність ентропії від обраного тексту та характеру символів, які він містить.