Національний технійчний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Фізико-технічний інститут

«Криптографія»

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

ФБ-21 Жиговець Олександр

ФБ-21 Альгішиєв Дмитро

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.

Варіант: 3

Хід роботи

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.

1. частоти букв (без пробілів)

	Символ, Кількість, Частота
2	о,54112,0.11450870
3	е,40034,0.08471764
4	а,37690,0.07975741
5	н,30532,0.06461006
6	и,30495,0.06453176
7	т,30441,0.06441749
8	с,24699,0.05226660
9	л,21911,0.04636680
10	в,21663,0.04584199
11	р,19953,0.04222339
12	к,15760,0.03335040
13	д,15283,0.03234100
14	м,15006,0.03175483
15	у,14294,0.03024814
16	п,13073,0.02766433
17	ь,11152,0.02359922
18	я,10109,0.02139208
19	ч,8427,0.01783273
20	б,8284,0.01753012
21	г,7808,0.01652284
22	ы,7781,0.01646570
23	з,7589,0.01605940
24	ж,5345,0.01131078
25	й,4630,0.00979774
26	х,4193,0.00887298
27	ш,4101,0.00867830
28	ю,2657,0.00562259
29	э,1546,0.00327156
30	щ,1445,0.00305783
31	ц,1389,0.00293932
32	ё,546,0.00115541
33	ф,497,0.00105172
34	ъ,113,0.00023912

2. частоти букв (з пробілами)

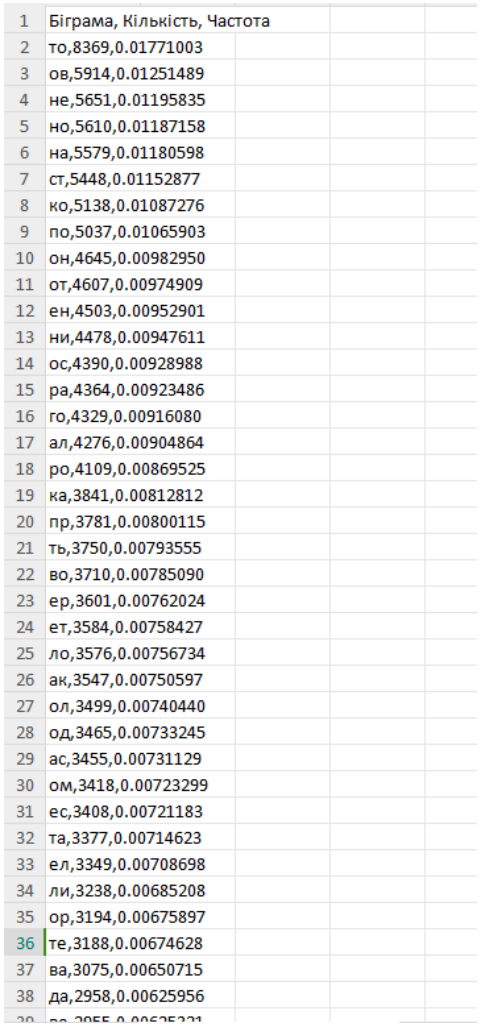
	Символ, Кількість, Частота
2	[space],96840,0.17007436
3	о,54112,0.09503370
4	е,40034,0.07030934
5	а,37690,0.06619272
6	н,30532,0.05362154
7	и,30495,0.05355656
8	т,30441,0.05346173
9	с,24699,0.04337739
10	л,21911,0.03848099
11	в,21663,0.03804544
12	р,19953,0.03504227
13	к,15760,0.02767836
14	д,15283,0.02684063
15	м,15006,0.02635415
16	у,14294,0.02510371
17	п,13073,0.02295934
18	ь,11152,0.01958560
19	я,10109,0.01775384
20	ч,8427,0.01479984
21	б,8284,0.01454870
22	г,7808,0.01371273
23	ы,7781,0.01366531
24	з,7589,0.01332811
25	ж,5345,0.00938711
26	й,4630,0.00813139
27	х,4193,0.00736392
28	ш,4101,0.00720234
29	ю,2657,0.00466633
30	э,1546,0.00271515
31	щ,1445,0.00253777
32	ц,1389,0.00243942
33	ё,546,0.00095891
34	ф,497,0.00087285
35	ъ,113,0.00019846

Частоти біграм (з перетинами, з пробілами)

	Біграма, Кількість, Частота
2	о[space],13740,0.02413079
3	е[space],10022,0.01760108
4	и[space],9992,0.01754839
5	а[space],9750,0.01712338
6	[space]в,9401,0.01651045
7	[space]н,9213,0.01618027
8	[space]п,9120,0.01601694
9	[space]с,8790,0.01543738
10	то,8150,0.01431339
11	ь[space],6936,0.01218131
12	[space]и,6681,0.01173347
13	[space]о,6549,0.01150164
14	я[space],6443,0.01131548
15	[space]т,5956,0.01046019
16	не,5599,0.00983321
17	на,5549,0.00974540
18	но,5441,0.00955572
19	ст,5300,0.00930809
20	по,5036,0.00884444
21	ко,4970,0.00872853
22	[space]д,4750,0.00834216
23	[space]к,4685,0.00822800
24	ов,4361,0.00765898
25	ра,4354,0.00764669
26	м[space],4346,0.00763264
27	ни,4304,0.00755887
28	го,4265,0.00749038
29	л[space],4228,0.00742540
30	ал,4090,0.00718304
31	ро,4079,0.00716372
32	у[space],4029,0.00707591
33	[space]б,3984,0.00699688
34	ка,3818,0.00670534
35	пр,3781,0.00664036
36	ть,3750,0.00658591
37	от,3701,0.00649986
38	ен,3637,0.00638746

bigrams_with_spaces_overlapping +

Частоти біграм (з перетинами, без пробілів)



	Біграма, Кількість, Частота
1	то,8369,0.01771003
2	ов,5914,0.01251489
3	не,5651,0.01195835
4	но,5610,0.01187158
5	на,5579,0.01180598
6	ст,5448,0.01152877
7	ко,5138,0.01087276
8	по,5037,0.01065903
9	он,4645,0.00982950
10	от,4607,0.00974909
11	ен,4503,0.00952901
12	ни,4478,0.00947611
13	ос,4390,0.00928988
14	ра,4364,0.00923486
15	го,4329,0.00916080
16	ал,4276,0.00904864
17	ро,4109,0.00869525
18	ка,3841,0.00812812
19	пр,3781,0.00800115
20	ть,3750,0.00793555
21	во,3710,0.00785090
22	ер,3601,0.00762024
23	ет,3584,0.00758427
24	ло,3576,0.00756734
25	ак,3547,0.00750597
26	ол,3499,0.00740440
27	од,3465,0.00733245
28	ас,3455,0.00731129
29	ом,3418,0.00723299
30	ес,3408,0.00721183
31	та,3377,0.00714623
32	ел,3349,0.00708698
33	ли,3238,0.00685208
34	ор,3194,0.00675897
35	те,3188,0.00674628
36	ва,3075,0.00650715
37	да,2958,0.00625956
38	ле,2955,0.00625321

Частоти біграм (без перетинів, з пробілами)

	Біграма, Кількість, Частота
1	о[space],6816,0.02394107
2	е[space],5014,0.01761158
3	и[space],5002,0.01756943
4	а[space],4829,0.01696177
5	[space]в,4690,0.01647354
6	[space]н,4591,0.01612580
7	[space]п,4580,0.01608717
8	[space]с,4396,0.01544087
9	то,4085,0.01434849
10	т[space],3466,0.01217426
11	[space]к,3377,0.01186165
12	[space]о,3361,0.01180545
13	я[space],3191,0.01120833
14	[space]т,2947,0.01035128
15	на,2807,0.00985954
16	не,2789,0.00979631
17	но,2730,0.00958907
18	ст,2642,0.00927998
19	по,2531,0.00889009
20	ко,2385,0.00837727
21	[space]н,2357,0.00827892
22	[space]д,2323,0.00815949
23	ра,2224,0.00781176
24	ов,2212,0.00776961
25	го,2143,0.00752725
26	ни,2142,0.00752374
27	м[space],2125,0.00746402
28	л[space],2107,0.00740080
29	ал,2048,0.00719356
30	у[space],2043,0.00717600
31	ро,1996,0.00701091
32	[space]б,1987,0.00697930
33	ка,1946,0.00683529
34	пр,1890,0.00663967
35	ть,1891,0.00664210
36	ен,1844,0.00647702
37	от,1818,0.00638569
38	ле,1784,0.00626688

bigrams_with_spaces_nonoverlapp

Частоти біграм (без перетинів, без пробілів)

	Біграма	Кількість	Частота
1	то	4202	0.01778406
2	ов	2950	0.01248524
3	но	2804	0.01186733
4	не	2791	0.01181231
5	ст	2750	0.01163878
6	на	2744	0.01161339
7	ко	2542	0.01075847
8	по	2480	0.01049607
9	от	2301	0.00973849
10	он	2293	0.00970463
11	ни	2251	0.00952687
12	ен	2244	0.00949725
13	ра	2221	0.00939990
14	ал	2207	0.00934065
15	ос	2154	0.00911634
16	го	2131	0.00901900
17	ро	2109	0.00892589
18	ка	1876	0.00793977
19	ть	1866	0.00789744
20	пр	1852	0.00783819
21	во	1851	0.00783396
22	ер	1827	0.00773238
23	ло	1800	0.00761811
24	ол	1792	0.00758425
25	ет	1759	0.00744459
26	од	1757	0.00743612
27	ак	1747	0.00739380
28	ом	1718	0.00727107
29	ел	1693	0.00716526
30	та	1689	0.00714833
31	ас	1664	0.00704252
32	ес	1657	0.00701290
33	те	1594	0.00674626
34	ва	1589	0.00672510
35	ор	1560	0.00660236
36	ли	1547	0.00654734
37	ат	1511	0.00639498

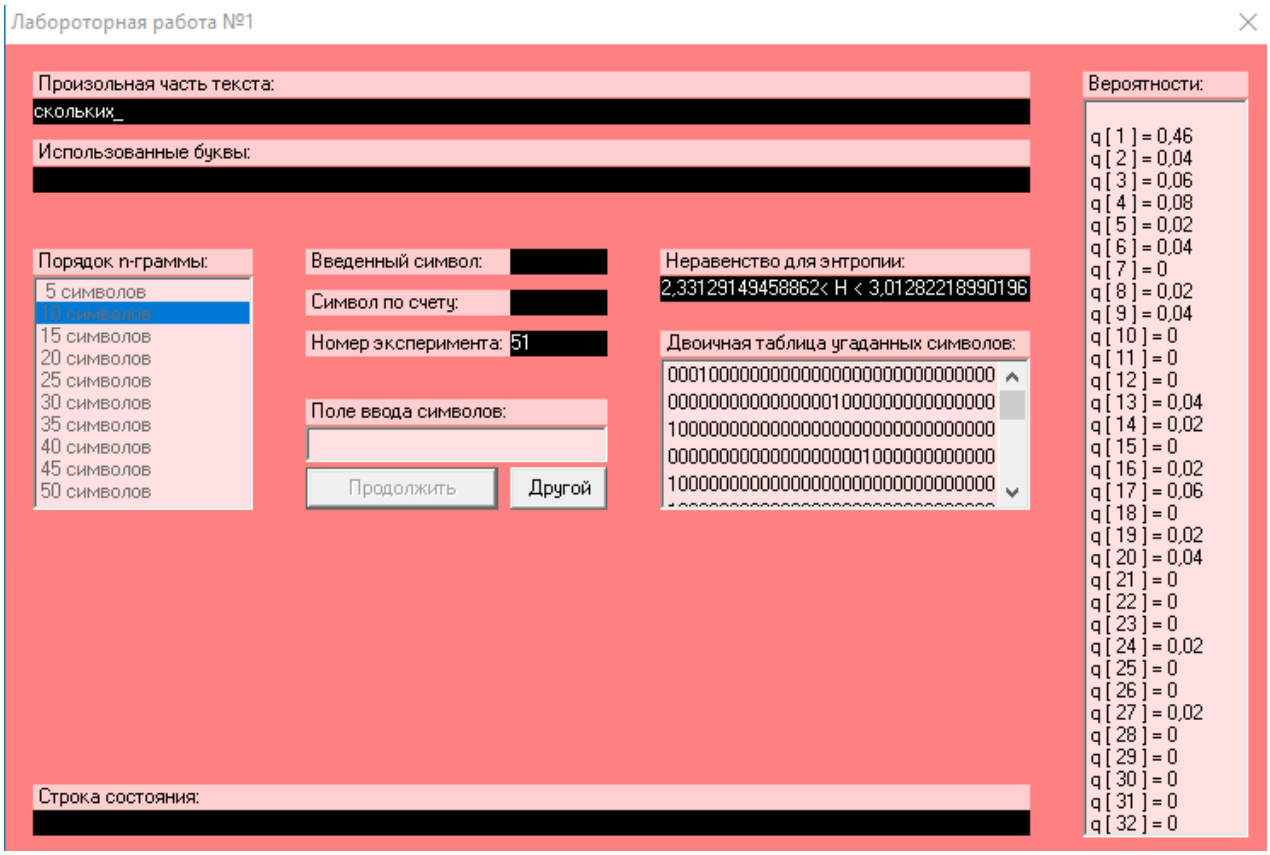
bigrams_no_spaces_nonoverlappin

Результати

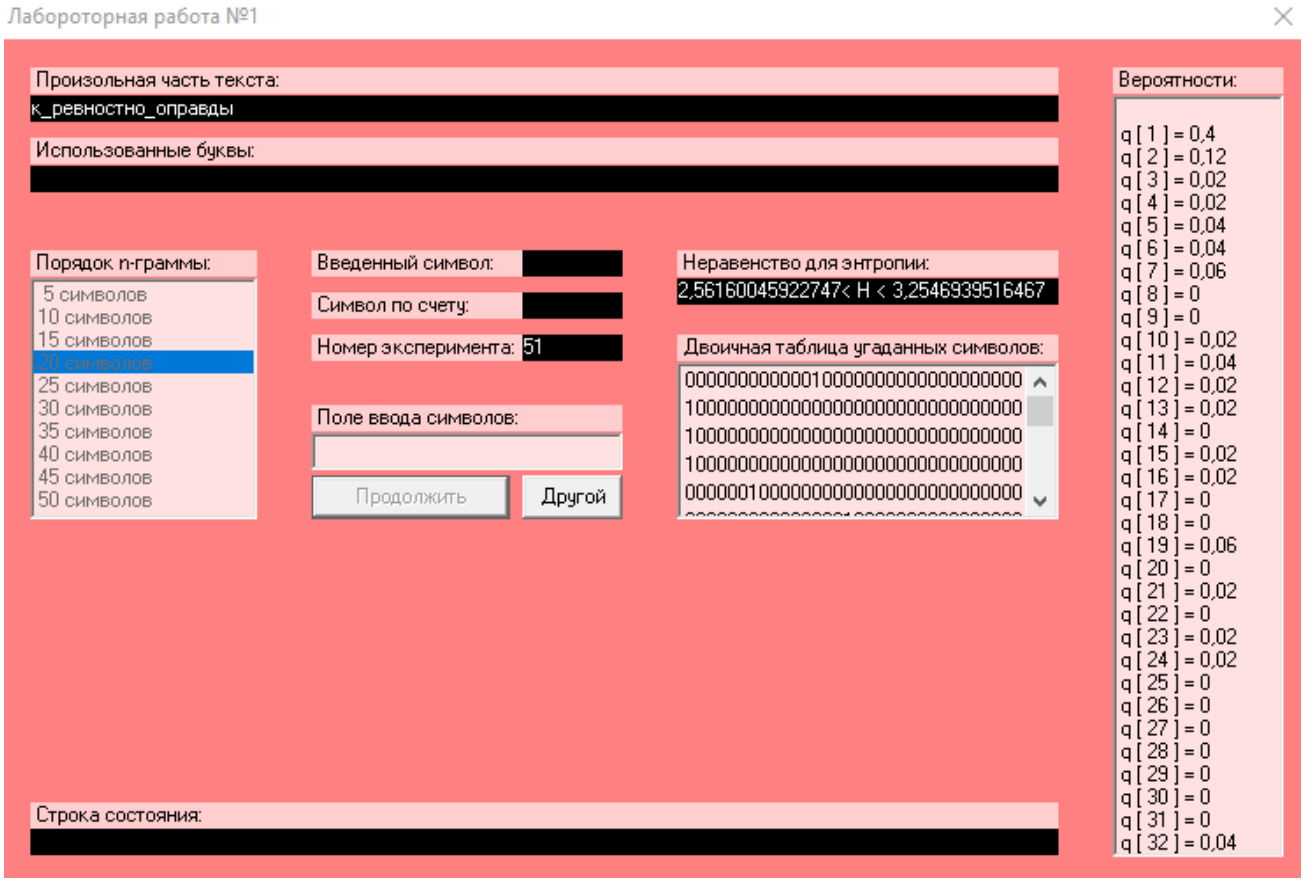
```
=== ТЕКСТ З ПРОБІЛАМИ ===
Довжина тексту з пробілами: 569398
Кількість унікальних символів: 34
H1 (на символ)   = 4.362443
R    = 0.142511

--- Біграми (перекривні) ---
H2 (перекривні)   = 3.946628
R  = 0.224244

--- Біграми (неперекривні) ---
H2 (неперекривні)  = 3.946911
R   = 0.224189

============================================================
=== ТЕКСТ БЕЗ ПРОБІЛІВ ===
Довжина тексту без пробілів: 472558
Кількість унікальних символів: 33
H1 (на символ)    = 4.463735
R    = 0.115110

--- Біграми (перекривні) ---
H2 (перекривні)    = 4.136764
R  = 0.179928

--- Біграми (неперекривні) ---
H2 (неперекривні)  = 4.137789
R  = 0.179725
```

2. За допомогою програми CoolPinkProgram оцінити значення H(10) , H(20) , H (30)

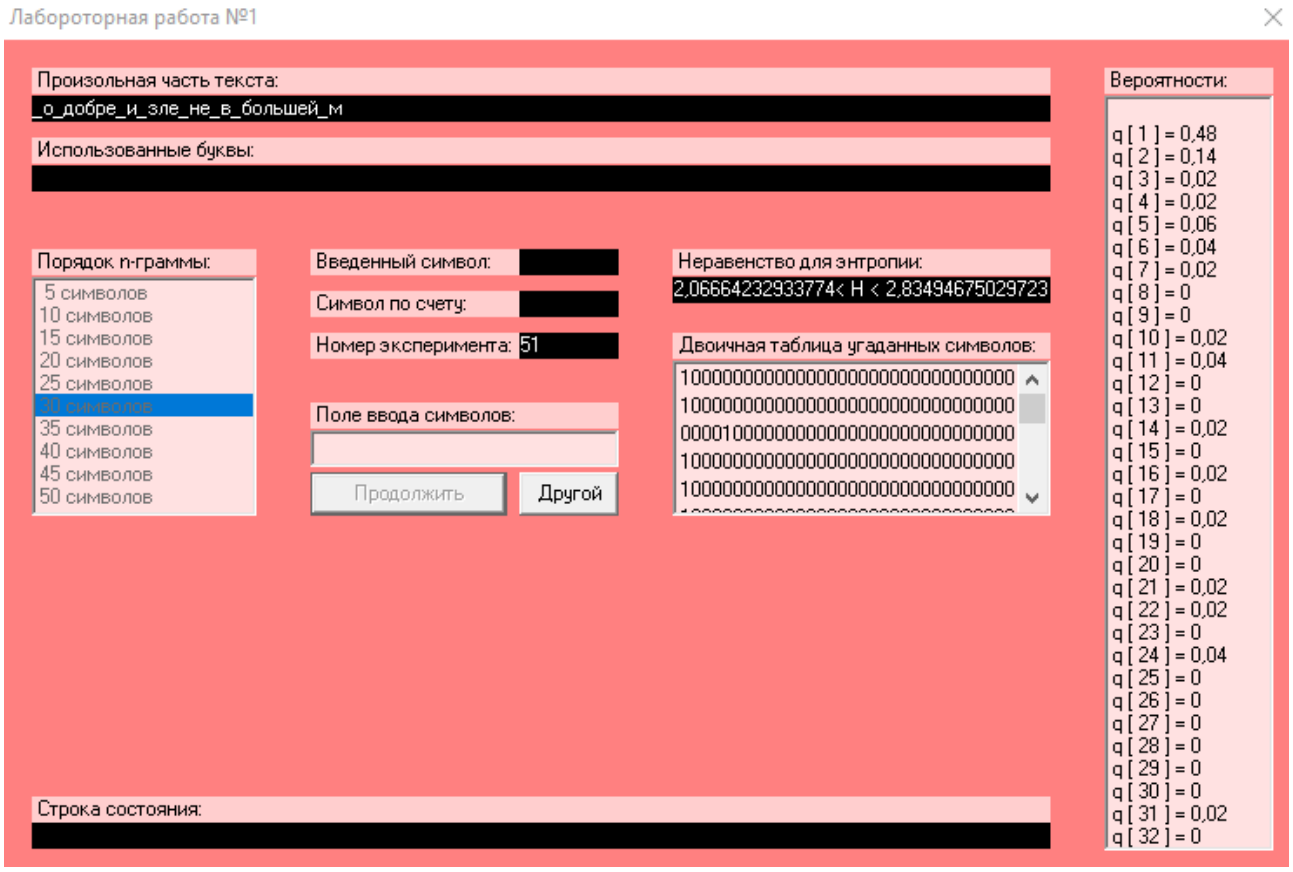
R = 1- H/log2(alphabet_size)



2,331 < H < 3,012
0,533 < R < 0,397



2,561 < H < 3,254
0,487 < R < 0,349



2,066 < H < 2,834

0,586 < R < 0,433

Висновки: При виконанні цієї лабораторної роботи ми засвоїли поняття ентропії та надлишковості, проаналізували частоти біграм та символів у тексті. Через це ми змогли зрозуміти як це впливає на кількість інформації що міститься в текстах