Міністерство освіти і науки України

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

Криптографія

Лабораторна робота №1

Виконав студент групи ФБ-13

Нійозов Рустам

Київ 2023

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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 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.

Аналіз результатів для тексту з пробілами

Текст без пробілів

а	0,081296
б	0,016728
в	0,046564
г	0,018979
д	0,031136
е	0,087972
ё	0
э	0,003554
ж	0,011568
з	0,017309
и	0,063784
ы	0,017417
й	0,010371
к	0,0328
л	0,046495
м	0,03076
н	0,067019
о	0,111932
п	0,027224
р	0,039487
с	0,053293
т	0,063036
у	0,026203
ф	0,002211
х	0,007079
ц	0,003261
ч	0,018446
ш	0,008342
щ	0,002955
ъ	0,00022
ь	0,022638
ю	0,005954
я	0,023967

Текст з пробілами

а	0,067572
б	0,013904
в	0,038703
г	0,015775
д	0,02588
е	0,073121
ё	0
э	0,002954
ж	0,009615
з	0,014387
и	0,053017
ы	0,014476
й	0,00862
к	0,027263
л	0,038646
м	0,025567
н	0,055705
о	0,093036
п	0,022628
р	0,032821
с	0,044296
т	0,052395
у	0,021779
ф	0,001838
х	0,005884
ц	0,002711
ч	0,015332
ш	0,006934
щ	0,002456
ъ	0,000183
ь	0,018817
ю	0,004949
я	0,019921

Біграми

Текст з пробілами

Текст без пробілів

аа	0
аб	0,00041
ав	0,00288
аг	0,00075
ад	0,00175
ае	0,00128
аё	0
аэ	0
аж	0,00172
аз	0,00324
аи	0,00022
ау	0
ай	0,00078
ак	0,00541
ал	0,00805
ам	0,00322
ан	0,00341
ао	1,1E-05
ап	0,00069
ар	0,0021
ас	0,00544
ат	0,00379
ау	6,6E-05
аф	0,00021
ах	0,0007
ац	2,6E-05
ач	0,00089
аш	0,0008
ащ	0,00017
аъ	0
аь	0
аю	0,00076
ая	0,00187
а	0,0166
ба	0,00062
бб	0
бв	2,6E-05
бг	1,1E-05
бд	7,3E-06
бе	0,00231
бё	0
бэ	0
бж	3,7E-06
бз	0
би	0,00071

аа	0,00067
аб	0,00112
ав	0,00573
аг	0,00156
ад	0,00306
ае	0,00212
аё	0
аэ	0,00039
аж	0,00219
аз	0,0042
аи	0,00165
ау	0
ай	0,00098
ак	0,00754
ал	0,0099
ам	0,00479
ан	0,00628
ао	0,00123
ап	0,00242
ар	0,00285
ас	0,00816
ат	0,00581
ау	0,00059
аф	0,00036
ах	0,00098
ац	7,3E-05
ач	0,00168
аш	0,00105
ащ	0,00023
аъ	0
аь	0
аю	0,001
ая	0,0027
ба	0,00071
бб	6,6E-06
бв	5,1E-05
бг	8,8E-06
бд	2,2E-05
бе	0,00283
бё	0
бэ	0,00011
бж	2,2E-06
бз	1,1E-05
би	0,00081
бы	0,00465

Перехрестні біграми

Текст з пробілами

аа	0
аб	0,00040786
ав	0,00303793
аг	0,00077732
ад	0,00173936
ае	0,00129492
аё	0
аэ	0
аж	0,00166803
аз	0,00313853
аи	0,00025057
аы	0
ай	0,00081572
ак	0,00534428
ал	0,00800727
ам	0,00327936
ан	0,00350067
ао	0,00000732
ап	0,00063648
ар	0,00202285
ас	0,00538817
ат	0,00381891
ау	0,00006036
аф	0,00025606
ах	0,00068953
ац	0,00002926
ач	0,0008962
аш	0,00081024
ащ	0,00018107
аъ	0
аь	0
аю	0,00083219
ая	0,00186373
а	0,01681747
ба	0,0005871
бб	0
бв	0,00003658
бг	0,00000732
бд	0,00001646
бе	0,00233195
бё	0
бэ	0
бж	0,00000183
бз	0
би	0,00064014

Текст без пробілів

аа	0,00066894
аб	0,00112443
ав	0,00573218
аг	0,00156012
ад	0,00306303
ае	0,00211684
аё	0
аэ	0,00038728
аж	0,00219385
аз	0,00419626
аи	0,00165474
аы	0
ай	0,0009814
ак	0,00753655
ал	0,00989763
ам	0,00479038
ан	0,00627569
ао	0,00123445
ап	0,0024161
ар	0,00284959
ас	0,00815708
ат	0,00580699
ау	0,00058752
аф	0,00035647
ах	0,0009792
ац	0,00007261
ач	0,00168115
аш	0,00104962
ащ	0,00022665
аъ	0
аь	0
аю	0,00100341
ая	0,00269555
ба	0,00071295
бб	0,0000066
бв	0,00005061
бг	0,0000088

Значення H і R

Ентропія

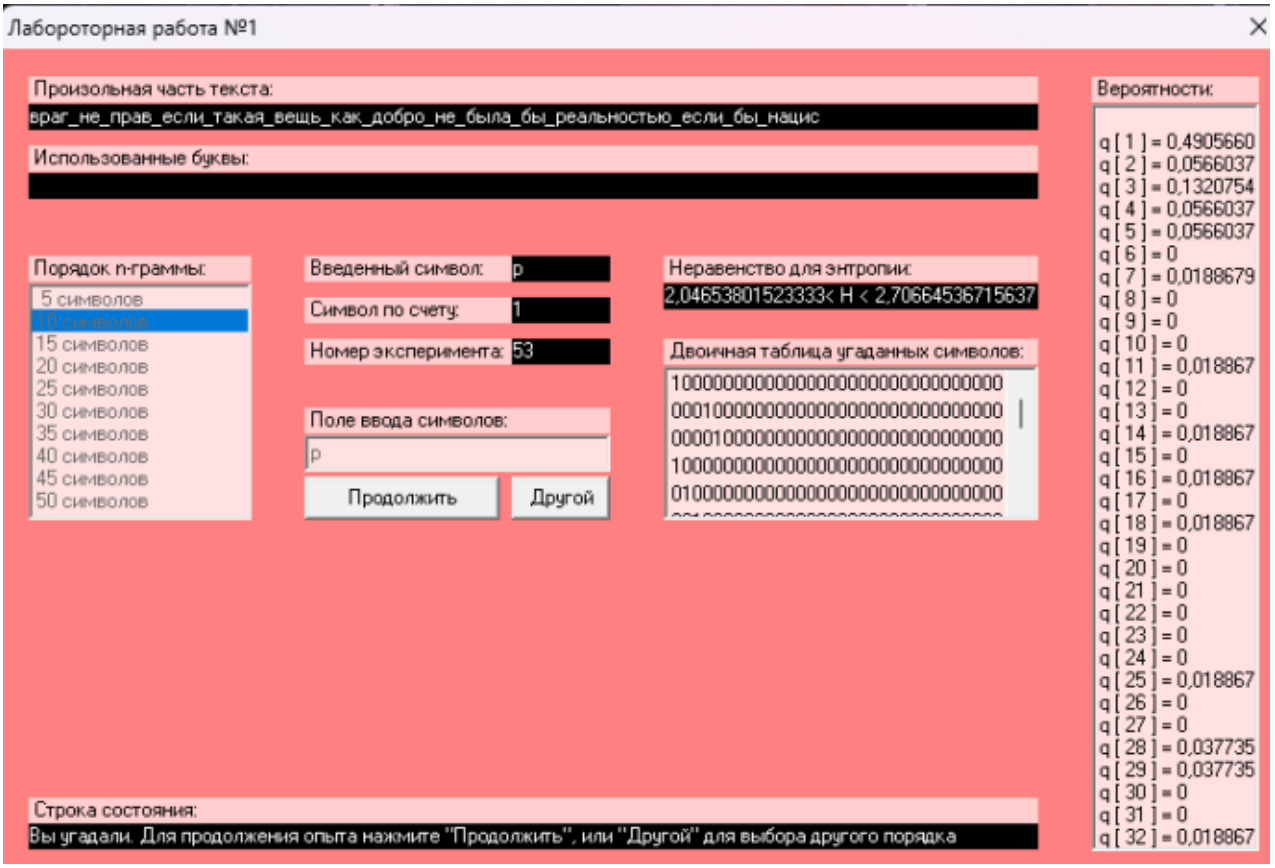
Ентропія на символ стаціонарного джерела визначається як

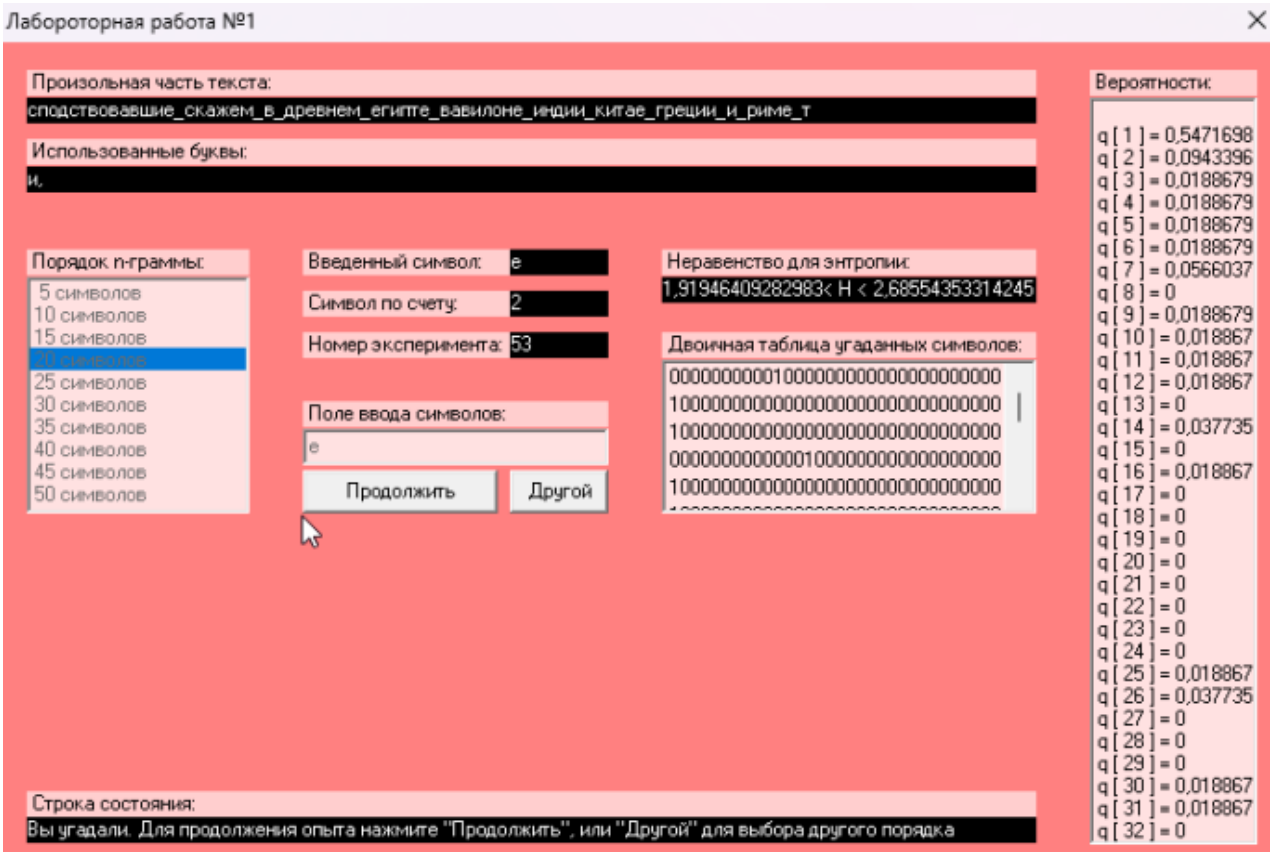
$$H_\infty = \lim_{n\to\infty} H_n, \text{ де } H_n = \frac{1}{n} H(x_1, x_2, ..., x_n),$$

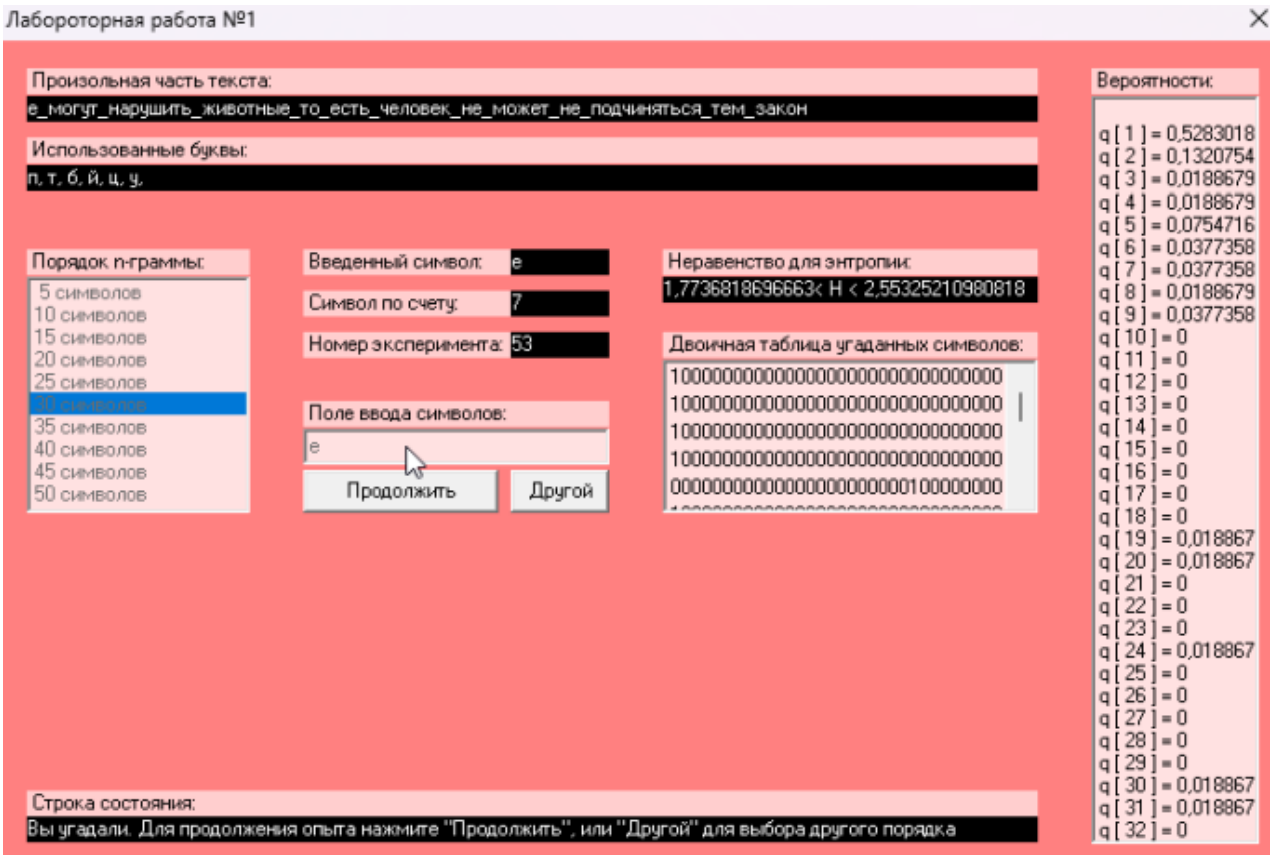
Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \frac{H_\infty}{H_0}$

```
letters with space: E =  4.363926750671057 , R =  0.1422194349436191
letters without space: E =  4.4622350062498795 , R =  0.11540714292613852
not crossed bigram with space: E =  3.9464790335576905 , R =  0.224273639591288
crossed bigram with space: E =  3.946960015449667 , R =  0.22417909700552663
not crossed bigram with space: E =  4.133355455939298 , R =  0.18060417997930378
crossed bigram with space: E =  4.134752310896859 , R =  0.1803272676436477
```

CoolPinkProgram







Отримані результати

$2,0465<H(10)<2,7066$

$1,9194<H(20)<2,6855$

1,7736<H(30)<2,5532
Висновки:Під час виконання лабораторної роботи, я отримав змогу ознайомитись з такими поняттями як ентропія, надлишковість та обрахувати їх на практиці. Успішно були проведені експерименти на різних видах тексту(з пробілом та без), прорахована частота монограм та біграм, проведено знайомство з невеличкою програмою. Набуті навички знадобляться у майбутніх лабораторних роботах та у професійній діяльності