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

Виконали:
Вісловух Владислав
Ісаченко Федір

Група: 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.

Порядок виконання 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. 2. За допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30). 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Під час виконання завдань, було використано текст двох книг. Всі маніпуляції з текстом відбуваються за допомогою функції open_sort().

Всі результати виконання роботи були виписані в окремі файли. Далі наведені таблиці з даними підписані та знаходяться нижче.

Частота літер з пробілом

```
1     ` : 0.160106
2     д : 0.025669
3     ж : 0.008671
4     о : 0.095158
5     р : 0.035819
6     у : 0.023955
7     э : 0.002714
8     л : 0.041818
9     п : 0.023468
10    е : 0.06951
11    в : 0.036845
12    а : 0.064994
13    я : 0.017197
14    ч : 0.012521
15    с : 0.044483
16    т : 0.054124
17    ь : 0.01648
18    б : 0.014714
19    ы : 0.017483
20    х : 0.00736
21    н : 0.056831
22    й : 0.009437
23    к : 0.028351
24    и : 0.057788
25    ц : 0.002844
26    г : 0.014005
27    з : 0.014325
28    м : 0.027693
29    ш : 0.006546
30    ю : 0.004707
31    щ : 0.003065
32    ф : 0.001149
33    ъ : 0.000171
```

Частота літер без пробілу

```
1     д : 0.030562
2     ж : 0.010324
3     о : 0.113297
4     р : 0.042647
5     у : 0.028522
6     э : 0.003232
7     л : 0.04979
8     п : 0.027941
9     е : 0.082761
10    в : 0.043869
11    а : 0.077384
12    я : 0.020475
13    ч : 0.014907
14    с : 0.052963
15    т : 0.064442
16    ь : 0.019621
17    б : 0.017519
18    ы : 0.020816
19    х : 0.008763
20    н : 0.067664
21    й : 0.011236
22    к : 0.033756
23    и : 0.068803
24    ц : 0.003386
25    г : 0.016675
26    з : 0.017056
27    м : 0.032972
28    ш : 0.007793
29    ю : 0.005604
30    щ : 0.003649
31    ф : 0.001368
32    ъ : 0.000203
```

Частота перехресних біграм з пробілами

`д : 0.00714	ьс : 0.000987
дж : 0.000552	ск : 0.003285
жо : 0.000169	ки : 0.002856
ор : 0.00462	ий : 0.00123
рд : 0.00031	де : 0.00466
ж` : 0.000173	ен : 0.007172
`о : 0.011106	нь : 0.001094
ру : 0.002578	`и : 0.009753
уэ : 9e-06	и` : 0.018101
эл : 2.9e-05	сы : 0.000357
лл : 0.000432	ы` : 0.004632
л` : 0.006383	ро : 0.006564
`п : 0.016626	об : 0.003791
пе : 0.001848	би : 0.000765
ер : 0.005389	ил : 0.004632
рв : 0.000421	ли : 0.007974
ва : 0.005642	`т : 0.007998
ая : 0.001736	тр : 0.002881
я` : 0.010535	ри : 0.004305
`ч : 0.005196	ин : 0.004543
ча : 0.001704	на : 0.009802
ас : 0.003537	ад : 0.001918
ст : 0.011832	дц : 0.000275
ть : 0.006034	ца : 0.000608
ь` : 0.010289	ат : 0.004577
`` : 0.00041	`у : 0.00478
`б : 0.006833	ут : 0.001739
бы : 0.003683	тк : 0.000662
ыл : 0.002336	кн : 0.000486
`х : 0.001223	ну : 0.002842
хо : 0.001762	ув : 0.000613
ол : 0.005897	в` : 0.005698
ло : 0.006583	по : 0.009329
од : 0.005024	дб : 3.7e-05
дн : 0.001899	бо : 0.002367
ны : 0.003763	до : 0.003634
ый : 0.001666	ок : 0.00216
й` : 0.007218	к` : 0.004236
`я : 0.002248	`в : 0.015195
яс : 0.000458	`г : 0.003351
сн : 0.001147	гр : 0.001042
`а : 0.002081	уд : 0.001906
ап : 0.000906	дь : 0.000449
пр : 0.006507	чт : 0.003263
ре : 0.005619	то : 0.014181
ел : 0.005587	`с : 0.015395
ль : 0.004034	сп : 0.001515

Частота перехресних біграм без пробілами

дж : 0.000663	ьи : 0.000939
жо : 0.000213	ич : 0.002042
ор : 0.00608	сы : 0.000425
рд : 0.000432	ып : 0.000887
ру : 0.003097	ро : 0.007871
уэ : 7.8e-05	об : 0.005881
эл : 3.5e-05	би : 0.000923
лл : 0.000608	ил : 0.005789
лп : 0.000647	ли : 0.009935
пе : 0.002201	ит : 0.005728
ер : 0.006986	тр : 0.003609
рв : 0.000593	ри : 0.005199
ва : 0.006791	ин : 0.007489
ая : 0.002391	на : 0.01172
яч : 0.000484	ад : 0.003083
ча : 0.002029	дц : 0.000327
ас : 0.005931	ца : 0.00073
ст : 0.014339	ат : 0.006397
ть : 0.007184	ьу : 0.000336
ьб : 0.000447	ут : 0.002351
бы : 0.004385	тк : 0.001107
ыл : 0.002846	кн : 0.001074
лх : 4.1e-05	ну : 0.003608
хо : 0.002381	ув : 0.00139
ол : 0.007543	вп : 0.000971
ло : 0.008652	по : 0.011119
од : 0.007052	дб : 0.000103
дн : 0.002461	бо : 0.002845
ны : 0.00448	до : 0.004377
ый : 0.001983	ок : 0.003724
йя : 8.4e-05	кв : 0.000609
яс : 0.001747	вг : 0.000332
сн : 0.001747	гр : 0.00126
йа : 0.000195	уд : 0.002532
ап : 0.002943	дь : 0.000535
пр : 0.007771	ьч : 0.000633
ре : 0.00671	чт : 0.003895
ел : 0.007094	то : 0.017308
ль : 0.004803	ыс : 0.001483
ьс : 0.002374	сп : 0.002163
ск : 0.004158	па : 0.002145
ки : 0.003721	ти : 0.005199
ий : 0.001466	ис : 0.00564
йд : 0.00056	сь : 0.003433
де : 0.005572	ьо : 0.000909
ен : 0.010468	от : 0.009335
нь : 0.001303	тз : 0.000171

Частота не перехресних Біграм з пробілами

```
`д : 0.007166481441504828
жо : 0.00013592062235654378
рд : 0.00029049701640908375
ж` : 0.000170567055506251
ор : 0.004522692081157937
уэ : 2.665110242285172e-06
лл : 0.000479719843611331
`п : 0.016646278573313186
ер : 0.005231611405605793
ва : 0.005663359264855991
я` : 0.010436571708788735
ча : 0.0017056705550625101
ст : 0.011825094145019309
ь` : 0.01021270244843678
`б : 0.006982588834787151
ыл : 0.0023159808005458146
`х : 0.0012046298295128979
ол : 0.00602847936804906
од : 0.004914463286773857
ны : 0.0037045032367763893
й` : 0.007094523464963128
яс : 0.0004743896231267606
ап : 0.0008848166004386772
ре : 0.005716661469701694
ль : 0.003936367827855199
ск : 0.0032780855980107616
ий : 0.0012605971446008864
ен : 0.007153155890293402
и` : 0.018224023836746008
сы : 0.0003437992212547872
ро : 0.006772045125646622
би : 0.0007835424112318406
ли : 0.007931368081040672
`т : 0.007798112568926414
ри : 0.004357455246136256
на : 0.009780954589186581
дц : 0.0002398599218056655
ат : 0.004544012963096218
ут : 0.0018069447442693467
кн : 0.00054101737918389
ув : 0.0006769380015404338
бо : 0.0023719481156338033
до : 0.0036432057012038303
к` : 0.00417889285990315
в` : 0.005706001028732553
гр : 0.0010420581047335023
уд : 0.0019242095949298943
```

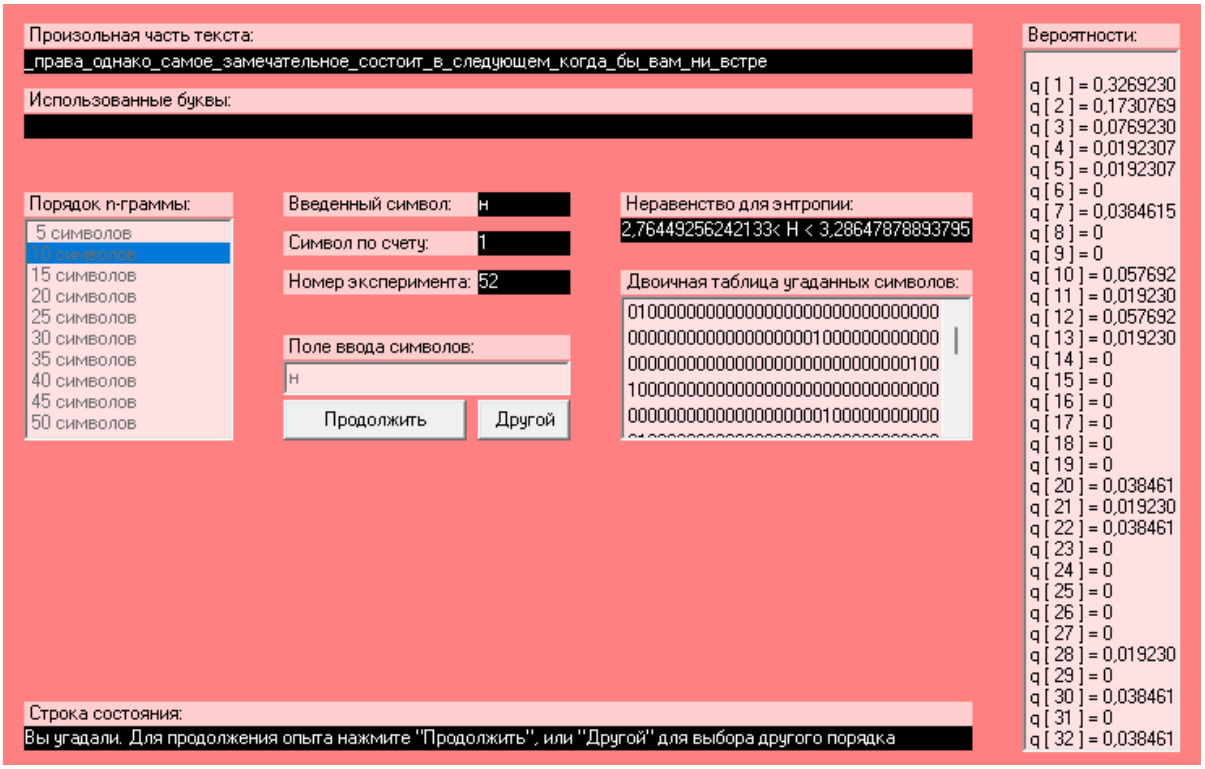
Частота не перехресних біграм без пробілами

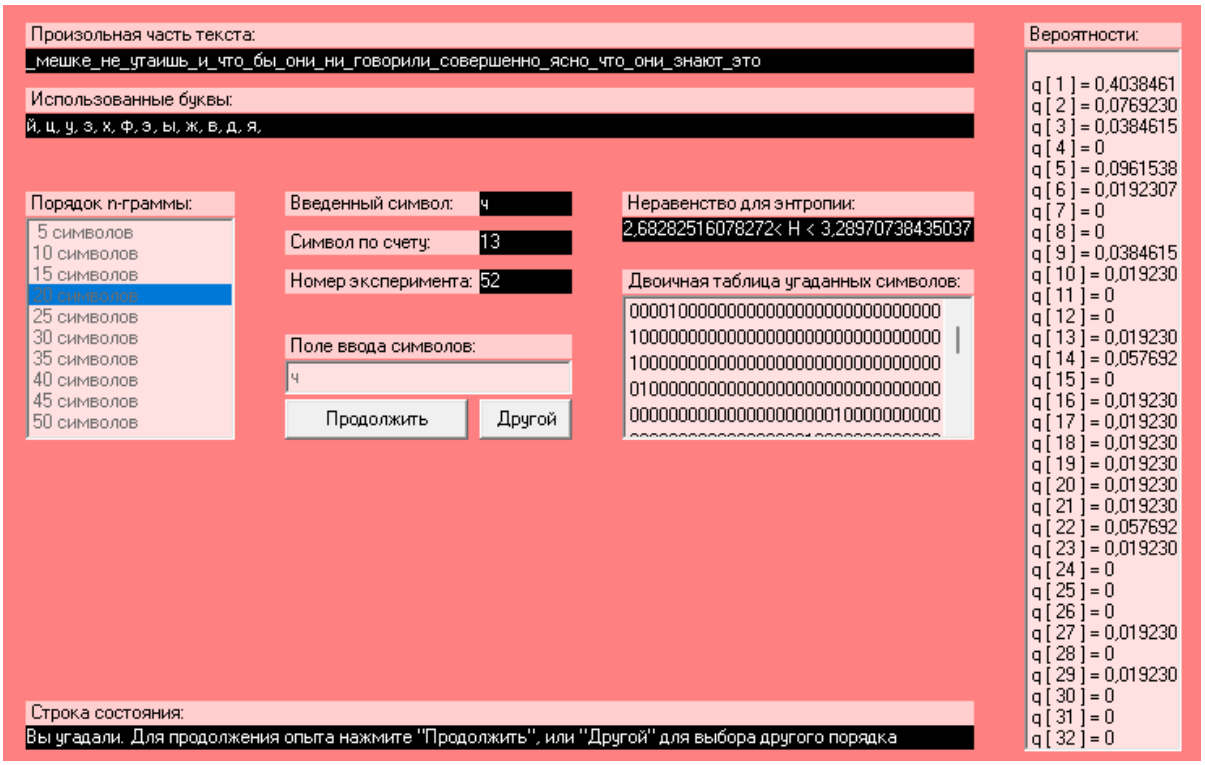
```
дж : 0.0006441499758840404
ор : 0.00600677785393344
уэ : 8.250196735460614e-05
лл : 0.0005933795344350519
пе : 0.002240245728936613
рв : 0.0005997258396161755
ая : 0.002427461731779758
ча : 0.001986393521691671
ст : 0.014431497981874953
ьб : 0.0004537608204503338
ыл : 0.00287804939963953
хо : 0.0023576523747873986
ло : 0.008494529484933871
дн : 0.0024052496636458253
ый : 0.001986393521691671
яс : 0.0017515802299900999
ны : 0.004623283324448506
йа : 0.00022529383392988603
пр : 0.007729799710608483
ел : 0.007149112786535679
ьс : 0.002497271088772117
ки : 0.0037855710405401977
йд : 0.0005553017033483106
ен : 0.010480923006625542
ьи : 0.0008980021831289823
сы : 0.00040299037900134543
об : 0.005832254461452542
ил : 0.0058735054451298455
ит : 0.005835427614043104
ри : 0.005203970248521311
на : 0.011575660650369355
дц : 0.0003268347168278628
ат : 0.006327266265580179
ьу : 0.00035856624273348055
тк : 0.0010947376437438123
ну : 0.00366499124209885
вп : 0.0009106947934912294
од : 0.007117381260630061
бо : 0.0027860279745132383
ро : 0.007828167440915898
до : 0.004182215114360419
кв : 0.0005616480085294342
гр : 0.0012248368999568451
уд : 0.002589292513898408
ьч : 0.0006155916025689844
то : 0.017255603787474934
бы : 0.004375777422384688
```

Ентропія та надлишковість

```
h1 with space:  4.37793521182762
R for h1 with space:  0.13946590895361488

h1 without space:  4.456959045060544
R for h1 without space:  0.11645304875040563

========================================================================

h2 with space с перетином:  3.983788720675063
R for h2 with space с перетином:  0.2169399865934014

h2 without space с перетином:  4.1459829534027355
R for h2 without space с перетином:  0.17810090660996525

h2 with space без перетину:  3.9828842143951113
R for h2 with space без перетину:  0.21711777782415353

h2 without space без перетину:  4.144479056738687
R for h2 without space без перетину:  0.17839903887886888
```

CoolPinkProgram





Результати

2.7644 < H(10) < 3.2864

2.6828 < H(20) < 3.2897

2.5846 < H(30) < 3.2472

H(10) H=3.0254 R=0.39492

H(20) H=2.9863 R=0.40274

H(30) H=2.9159 R=0.41682

Висновок

Під час роботи ми навчилися рахувати ентропію, та надлишковість російської мови на прикладі вибраного тексту. При виконанні робіт використовується Pyhon3 та програма CoolPinkProgram. Ми отримали знання і закріпили їх на практиці, ці знання будуть нами використовуватися у подальшому житті на роботі чи для особистих дій.