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:
Кравченко Владислав
Жмур Назар

Група: ФБ-04

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 роботи для читання файлу використовували функцію open()

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

Частота монограм з та без пробілів

texts > monogram_space.txt

```
 1    : 0.16464678482139297
 2   о : 0.08966730411162493
 3   е : 0.07288745772816296
 4   а : 0.07119721399468072
 5   и : 0.05583136187211497
 6   н : 0.05511134994888545
 7   т : 0.053850594377189684
 8   с : 0.044238959992107155
 9   л : 0.04389805638763931
10   в : 0.03445519447668988
11   р : 0.03420329526156585
12   к : 0.02907084875341376
13   м : 0.026181984587966355
14   д : 0.026017830266110528
15   у : 0.02373058539278435
16   п : 0.02180313656506032
17   ь : 0.017001097860745915
18   я : 0.016582945163640027
19   ы : 0.01601155377733369
20   б : 0.014641222047058972
21   г : 0.014563133290370523
22   з : 0.01334813940942229
23   ч : 0.012256576143884831
24   ж : 0.009779147363140008
25   й : 0.009067532080414626
26   х : 0.007816852477323822
27   ш : 0.007446140799066293
28   ю : 0.004786504919381755
29   э : 0.003830967230011272
30   щ : 0.002625629485642794
31   ц : 0.002553838209332446
32   ф : 0.000896761205841542
33
```

texts > monogram_no_space.txt

```
 1   о : 0.10734058657145787
 2   е : 0.08725345925984002
 3   а : 0.08523007118546619
 4   и : 0.06683563414570405
 5   н : 0.06597370902211956
 6   т : 0.06446446054041449
 7   с : 0.05295838836587038
 8   л : 0.052550293205316095
 9   в : 0.04124625829006118
10   р : 0.04094471014186835
11   к : 0.03480066662243961
12   м : 0.03134241194291493
13   д : 0.031145903066342604
14   у : 0.028407845880751778
15   п : 0.026100499966829702
16   ь : 0.02035198710178054
17   я : 0.019851417175780456
18   ы : 0.01916740545962974
19   б : 0.01752698353346079
20   г : 0.017433503607521012
21   з : 0.015979036372737634
22   ч : 0.01467232773056874
23   ж : 0.011706601693092336
24   й : 0.010854728174447614
25   х : 0.009357541618670253
26   ш : 0.008913763260579816
27   ю : 0.00529917395910605
28   э : 0.004586044753765834
29   щ : 0.003143136864663181
30   ц : 0.0030571956424282265
31   ф : 0.0010735114075664462
32
```

Частота неперехресних біграм з та без пробілів

texts > bigram_space.txt

```
а  : 0.02057681585823778
и  : 0.02044246955376875
о  : 0.0196825732691158
е  : 0.01926777905406767
 с : 0.01629704639649624
 н : 0.01626849780679658
 п : 0.01517189609656812
 в : 0.01499808556516131
то : 0.01199292666706970
 о : 0.01130020353465127
ь  : 0.01104074723414545
на : 0.01094418582780834
не : 0.01070824013058460
 и : 0.01002559297100135
ст : 0.00981315787705969
я  : 0.00972835177236362
ла : 0.00943111057372589
по : 0.00911035877180608
но : 0.00873167012608400
 к : 0.00851587637453062
ка : 0.00831519658223001
 д : 0.00790963867561413
ал : 0.00788780740113791
ли : 0.00779460465241252
ни : 0.00765438069712297
он : 0.00752423271466860
ко : 0.00737393278654387
 м : 0.00730172164789177
т  : 0.00728240936662434
 т : 0.00715897869939342
м  : 0.00711111782842633
й  : 0.00701539608649215
 б : 0.00690288105649933
ра : 0.00679120569090945
ет : 0.00669128562696061
го : 0.00667785099651371
ен : 0.00663334878315834
ть : 0.00623702718497470
у  : 0.00613962611423466
ро : 0.00613122947020534
от : 0.00605817866715031
```

texts > bigram_no_space.txt

```
то : 0.015251300175098959
на : 0.013123375409351611
не : 0.012645924174712976
ст : 0.011793548075821267
он : 0.011453803828857353
ла : 0.011396509680700717
по : 0.010862769458399422
но : 0.010462715581796945
ен : 0.01013402810026677
ка : 0.010118950692857129
ал : 0.009715881334772723
ли : 0.009683716198965488
ни : 0.009305775853230485
ов : 0.009259538470507585
ко : 0.009231393976676256
ос : 0.008985129655652117
ет : 0.008931856149471387
от : 0.00827548634690501
ра : 0.008142805161700168
го : 0.00799404140859171
ть : 0.007736720322133835
ас : 0.0077236532357121464
ер : 0.007604039136928993
ло : 0.0074693476307362
во : 0.007416074124555467
ро : 0.0073818986677602816
та : 0.0072833929393506264
ес : 0.007222078149218086
ол : 0.006978829309675876
ан : 0.006967772544242139
ле : 0.00695470545782045
пр : 0.006930581605965025
ат : 0.0069205300010252635
ом : 0.006894395828181886
ак : 0.006804936544218015
ел : 0.006728544346675833
ва : 0.006395836223169754
ре : 0.006250087951543223
од : 0.006219933136723941
ор : 0.006208876371290204
ит : 0.006119417087326333
```

Частота перехресних біграм з та без пробілів

texts > cross_bigram_space.txt

```
а  : 0.02070780350509508
и  : 0.02042567626571012
о  : 0.01967039813527329
е  : 0.01941094183476748
 с : 0.01640956142648906
 н : 0.01640284411126561
 п : 0.01520842149809564
 в : 0.01498549059911734
то : 0.01212727297153873
 о : 0.01124730467726659
ь  : 0.01088121099758848
на : 0.01087827217217822
не : 0.01054030724999832
 и : 0.01002265414559109
ст : 0.00969308586744050
я  : 0.00958980714587993
ла : 0.00957175436121690
по : 0.00909986296676944
но : 0.00865400116881284
 к : 0.00840336134453781
ка : 0.00834542450073554
 д : 0.00785967864363971
ли : 0.00782567223532098
ал : 0.00778620800838320
ни : 0.00759266536350751
он : 0.00755865895518879
ко : 0.00751373690963195
т  : 0.00732187359356212
 м : 0.00725218144811881
 т : 0.00713672759271574
й  : 0.00703764719316983
м  : 0.00702967038134198
 б : 0.00687727129220993
ра : 0.00683654756866775
ен : 0.00672277304207054
ет : 0.00671605572684709
го : 0.00671059790822803
ть : 0.00639488409272581
ро : 0.00613626745662293
у  : 0.00607959010942506
от : 0.00602921024524917
```

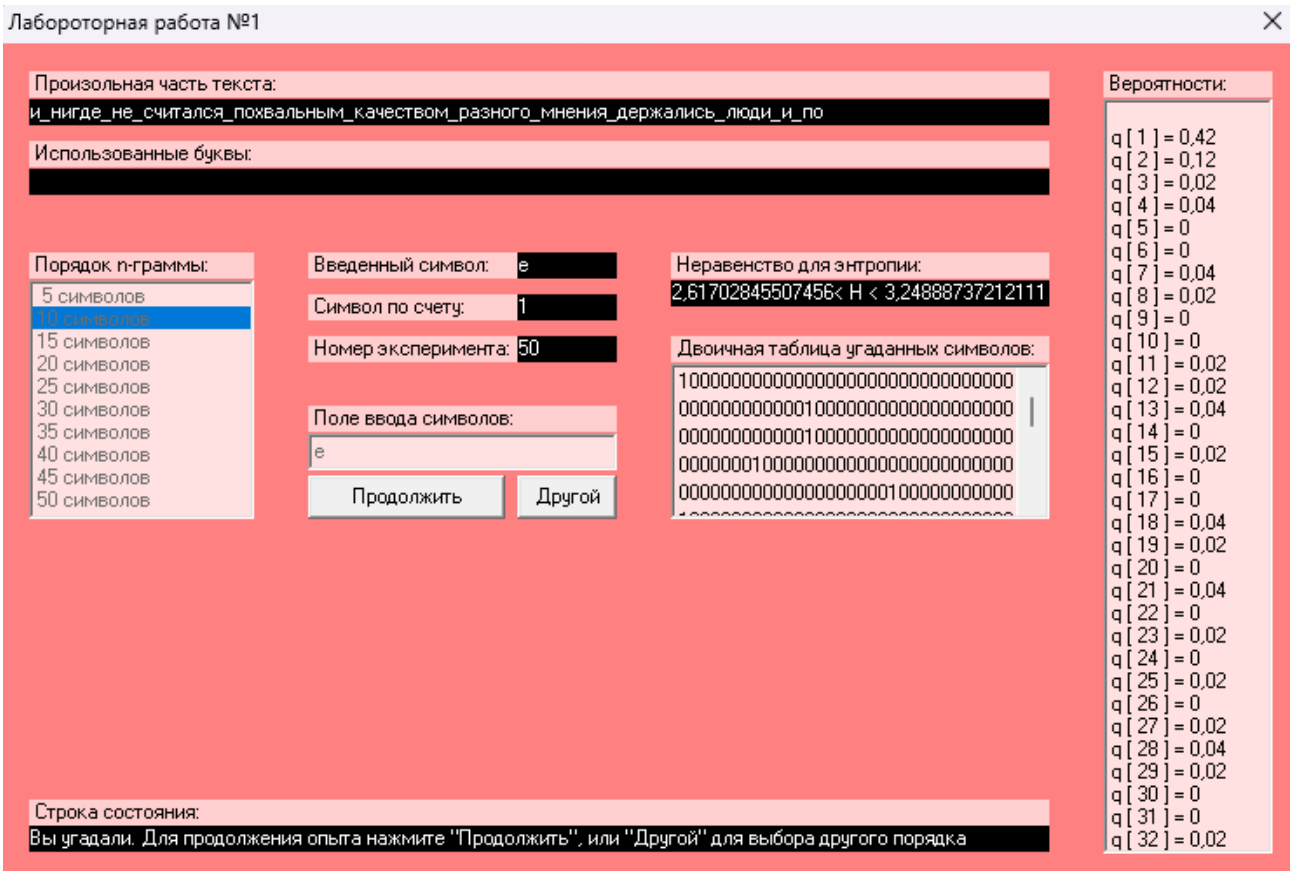
texts > cross_bigram_no_space.txt

```
то : 0.015110082719724425
на : 0.013064077505954322
не : 0.012698701482763248
ст : 0.011924727513417643
ла : 0.011511103762267362
он : 0.011327661879922462
по : 0.010893934908789179
но : 0.010561729198570059
ен : 0.010111417070950797
ка : 0.010043066123008587
ал : 0.009704326866294992
ли : 0.009615370117870205
ни : 0.009379659863569497
ко : 0.009359556643586495
ов : 0.009105753491301086
ос : 0.009045946411851653
ет : 0.008935378701945136
от : 0.008346354356443158
ра : 0.008198595689568088
го : 0.008087525399161997
ть : 0.007655306169527439
ас : 0.007620628115056759
ер : 0.007607561022067807
во : 0.007466838482186788
ро : 0.007416580432229281
ло : 0.00740451850023948
та : 0.00733013658630237
ес : 0.007209517266404353
ан : 0.007033614091553079
ом : 0.006951693470122343
ол : 0.006950688309123193
ле : 0.0069114870301563375
ат : 0.006909979288657613
пр : 0.00685519801420393
ак : 0.00677780061726937
ел : 0.006701408381333959
ва : 0.006461677483036651
ре : 0.0062787381811913266
од : 0.0062350136777282955
ит : 0.006142036285306908
ор : 0.006137010480311158
```

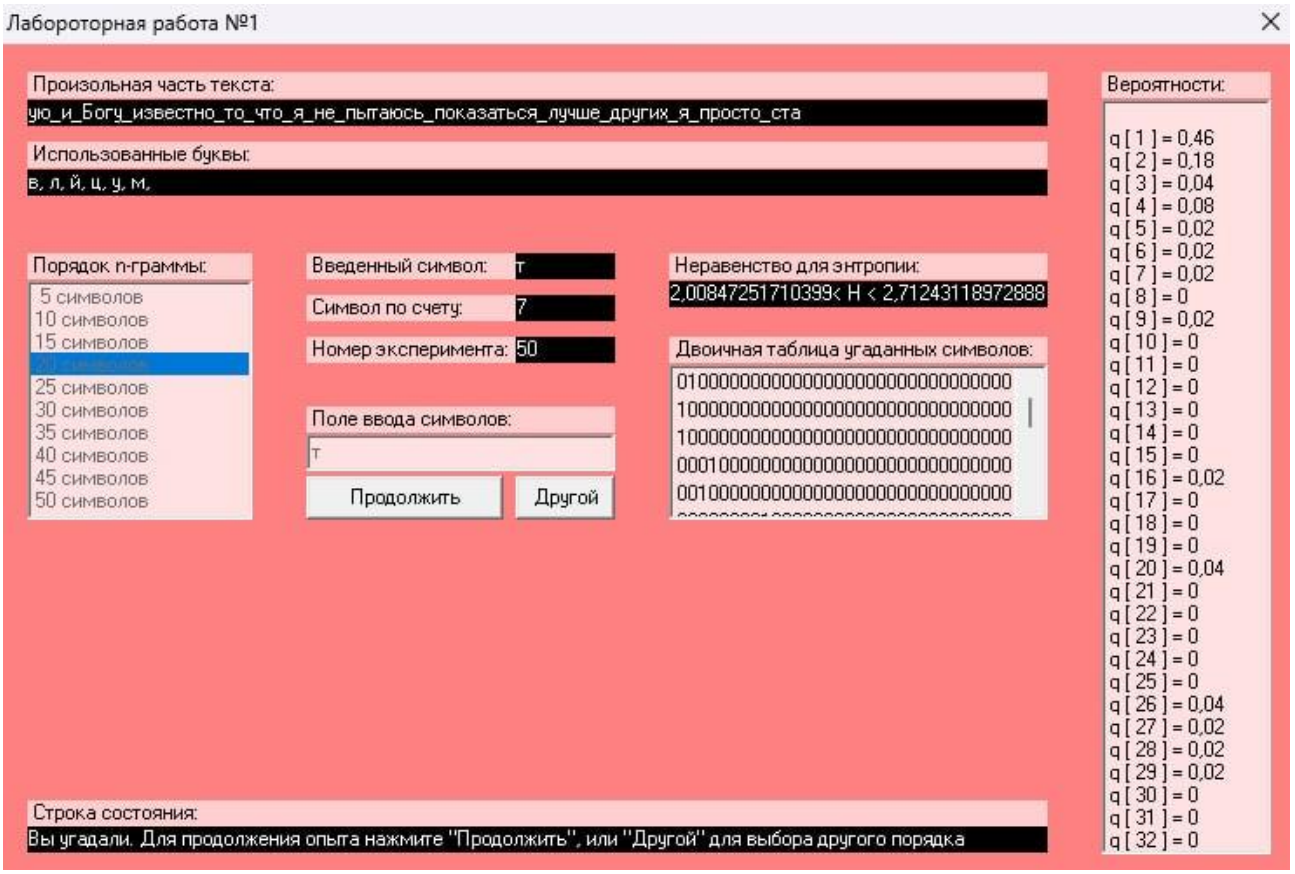
Ентропія та надлишковість

```
Монограми:

    Ентропія з пробілами: 4.3673644204588555
    Надлишковість з пробілами: 0.12652711590822885

    Ентропія без пробілів: 4.455664035773609
    Надлишковість без пробілів: 0.10886719284527824

Біграми без перетину:

    Ентропія з пробілами: 3.9694337769074917
    Надлишковість з пробілами: 0.20611324461850167

    Ентропія без пробілів: 4.14573870702101
    Надлишковість без пробілів: 0.1708522585957979

Біграми з перетином:

    Ентропія з пробілами: 3.9694757965555487
    Надлишковість з пробілами: 0.20610484068889023

    Ентропія без пробілів: 4.145451584049106
    Надлишковість без пробілів: 0.17090968319017874
```

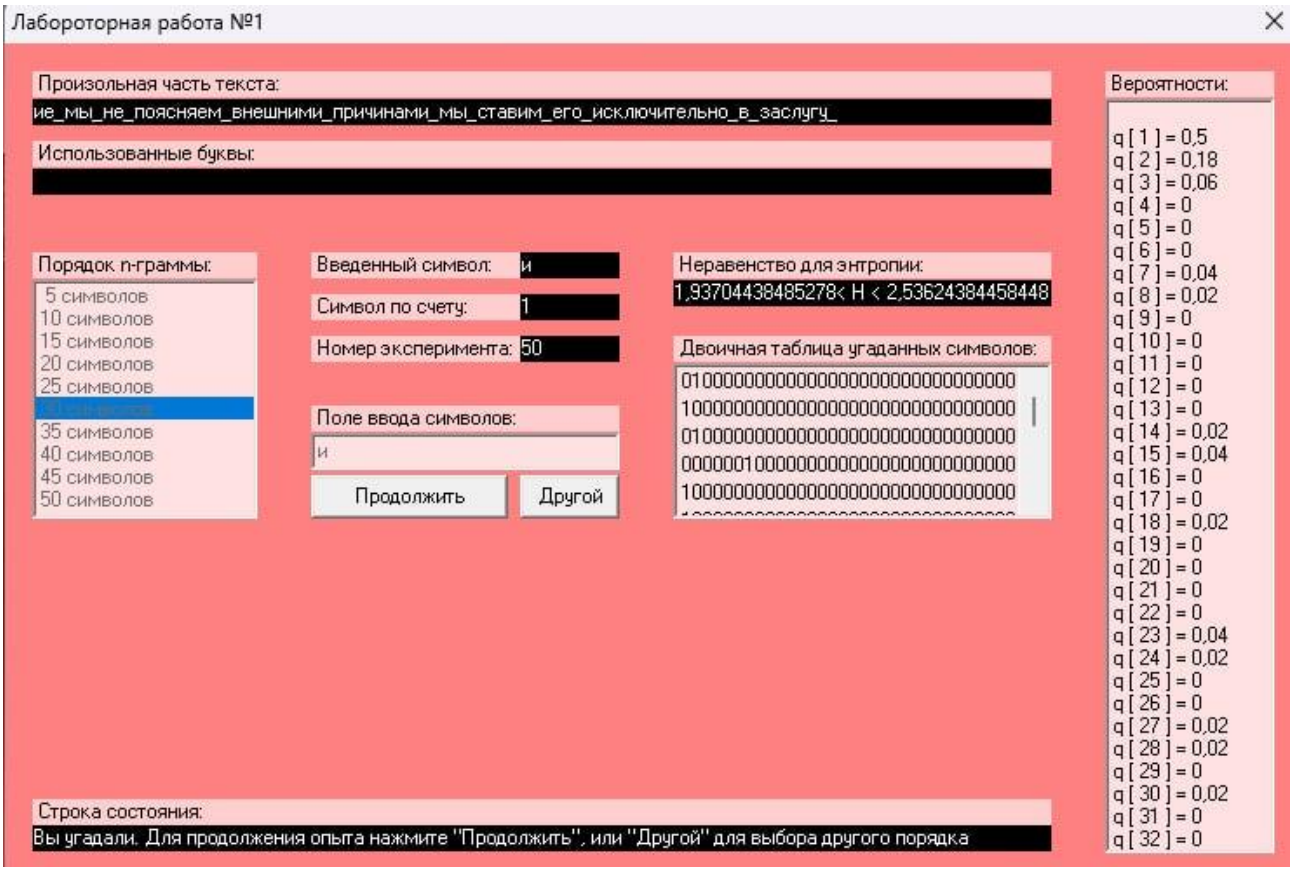
coolpinkprogram



R = 0.41341



R = 0.52792



R = 0.55268

Висновок

Під час роботи ми навчилися рахувати ентропію, та надлишковість російської мови на прикладі вибраного тексту. З отриамних даних можемо зробити висновок що найчастіше у нашому тексті зустрічаються букви «о», «е» та «а»