Міністерсво освіти і науки України
Національний технічний університет України
Київський політехнічний інститут імені Ігоря Сікорського
Навчально-науковий фізико-технічний інститут

Симетрична криптографія
Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
Медведцький Костянтин ФІ-04
Сковрон Роман ФІ-04

Київ 2023

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

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 H(10) , H(20) , H(30).
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Для роботи було обрано твір "Війна і мир"

1.1 Кількість та частота літер, якщо вважати, що пробіл не входить до алфавіту:

```
Monograms without spaces:
о: 265208
е: 189292
а: 188236
и: 155955
н: 148899
т: 134701
с: 122484
л: 115273
в: 106328
р: 101767
к: 77272
д: 70457
м: 68117
у: 62557
п: 60520
я: 50591
ь: 45777
г: 45263
ы: 44514
б: 39977
з: 39387
ч: 33218
й: 25743
ж: 24497
ш: 21319
х: 19779
ю: 14108
ц: 8280
э: 7050
щ: 6809
ф: 4707
```

```
Monogram frequency without spaces:
о: 11.540391238792298 %
е: 8.236945108644807 %
а: 8.190993805712148 %
и: 6.786302508392858 %
н: 6.479264256979181 %
т: 5.861445507890265 %
с: 5.329828966291499 %
л: 5.016045968708729 %
в: 4.6268088430149446 %
р: 4.4283392476779575 %
к: 3.3624517805042027 %
д: 3.0659005215211796 %
м: 2.9640766116135824 %
у: 2.722136039354506 %
п: 2.6334970203452004 %
я: 2.2014416351005295 %
ь: 1.9919628734359258 %
г: 1.9695964248493854 %
ы: 1.9370040707806717 %
б: 1.7395788232376086 %
з: 1.7139052733036417 %
ч: 1.4454643757737422 %
й: 1.1201935524578073 %
ж: 1.0659744961565827 %
ш: 0.92768544244447 %
х: 0.8606731256676754 %
ю: 0.6139024448617001 %
ц: 0.36029998890380466 %
э: 0.30677716446519604 %
щ: 0.2962901720345418 %
ф: 0.20482271108335856 %
```

1.2 Кількість та частота літер, якщо вважати, що пробіл входить до алфавіту:

```
Monograms with spaces:
 : 443926
о: 265208
е: 189292
а: 188236
и: 155955
н: 148899
т: 134701
с: 122484
л: 115273
в: 106328
р: 101767
к: 77272
д: 70457
м: 68117
у: 62557
п: 60520
я: 50591
ь: 45777
г: 45263
ы: 44514
б: 39977
з: 39387
ч: 33218
й: 25743
ж: 24497
ш: 21319
х: 19779
ю: 14108
ц: 8280
э: 7050
щ: 6809
ф: 4707
```

```
Monogram frequency with spaces:
 : 16.189796466899658 %
о: 9.672025385747904 %
е: 6.903400460464966 %
а: 6.864888579950993 %
и: 5.687613944655948 %
н: 5.430284561221672 %
т: 4.912489410144598 %
с: 4.466940504615043 %
л: 4.203958335688661 %
в: 3.8777379084183106 %
р: 3.711400136615061 %
к: 2.8180776809429284 %
д: 2.5695374672092854 %
м: 2.4841986410703676 %
у: 2.281427755030888 %
п: 2.2071392127894454 %
я: 1.845032715040166 %
ь: 1.6694681385304433 %
г: 1.6507227724469375 %
ы: 1.6234070541657202 %
б: 1.4579445523741517 %
з: 1.436427497920322 %
ч: 1.2114466353344315 %
й: 0.9388364962795555 %
ж: 0.8933953948397726 %
ш: 0.7774950574596527 %
х: 0.7213318983767752 %
ю: 0.5145128885332699 %
ц: 0.30196815403001664 %
э: 0.25711056593135473 %
щ: 0.24832139623072264 %
ф: 0.17166233104097686 %
```

1.3 Кількість та частота біграм без перетинів та без пробілів:

```
BiGram with no crossing and no spaces:
то: 20216
ст: 15511
на: 14404
ов: 13299
ал: 12369
ос: 12203
го: 12131
не: 12110
но: 12040
он: 11915
ен: 11781
по: 11644
ко: 11570
ра: 11163
от: 11013
ни: 11010
во: 10686
ка: 9741
ол: 9526
ло: 9404
ер: 9262
ла: 9218
ли: 9174
ор: 9072
ро: 9014
ел: 8673
ан: 8663
пр: 8496
ре: 7911
од: 7637
та: 7599
```

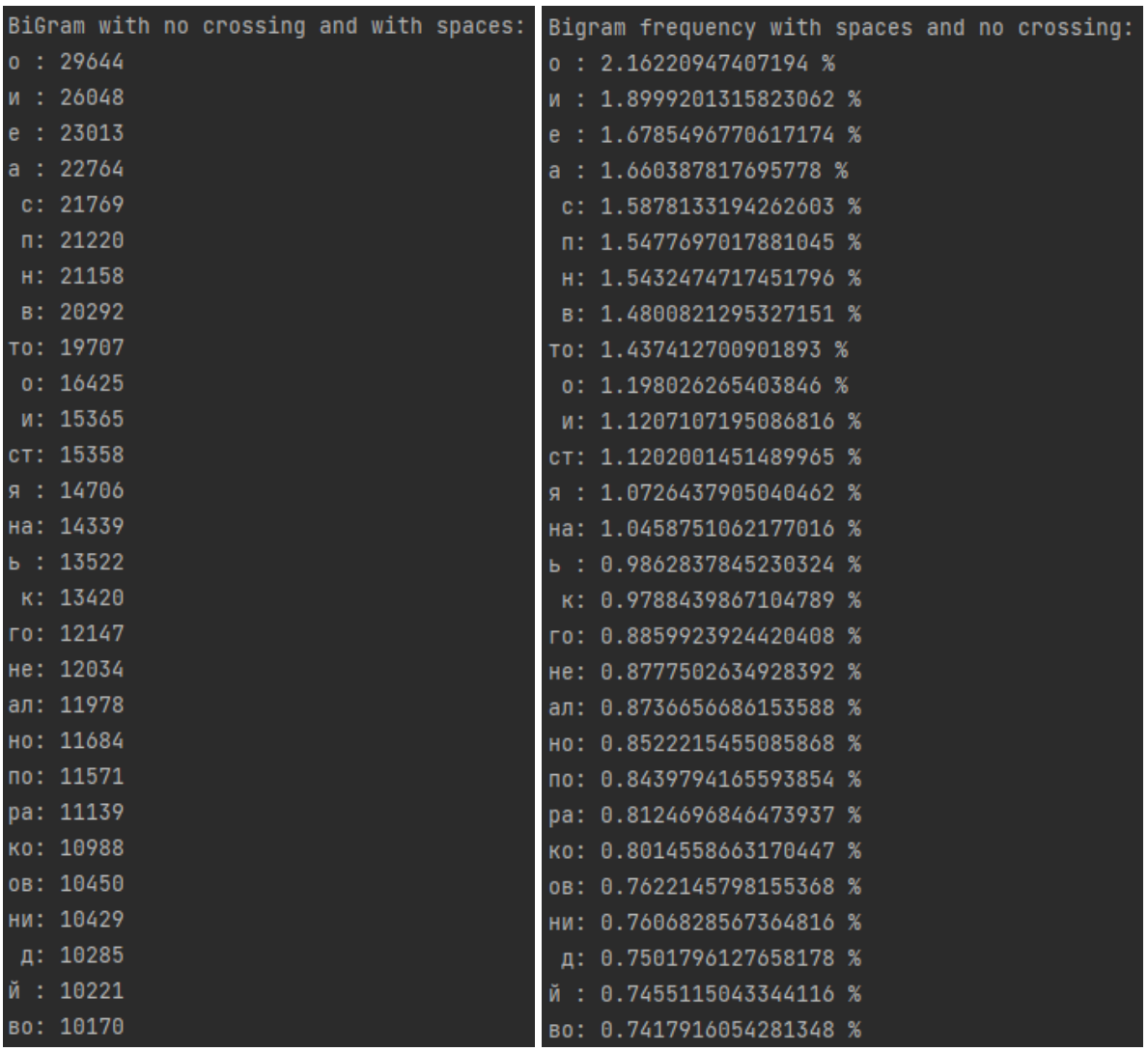
```
Bigram frequency without spaces and crossing:
то: 1.7593786824154383 %
ст: 1.349907140034916 %
на: 1.2535660141230693 %
ов: 1.1573989462526173 %
ал: 1.0764619570041827 %
ос: 1.0620151395684405 %
го: 1.055749050078239 %
не: 1.053921440643597 %
но: 1.0478294091947902 %
он: 1.0369507816076349 %
ен: 1.0252888928342045 %
по: 1.0133659169986824 %
ко: 1.0069257694670863 %
ра: 0.971504958043309 %
от: 0.9584506049387228 %
ни: 0.9581895178766311 %
во: 0.9299921151707249 %
ка: 0.8477496906118314 %
ол: 0.8290384511619244 %
ло: 0.818420910636861 %
ер: 0.8060627896978527 %
ла: 0.8022335127871739 %
ли: 0.7984042358764953 %
ор: 0.7895272757653768 %
ро: 0.7844795925649368 %
ел: 0.7548026965071772 %
ан: 0.7539324063002049 %
пр: 0.7393985598437655 %
ре: 0.6884865827358791 %
од: 0.6646406310648348 %
та: 0.6613335282783397 %
```

1.4 Кількість та частота біграм з перетинами, але без пробілів:

```
BiGram with crossing and without spaces:
то: 40383
ст: 31163
на: 28685
ов: 26796
ал: 24581
го: 24389
не: 24252
ос: 24107
но: 24062
ен: 23647
он: 23634
по: 23274
ко: 23103
ра: 22360
ни: 21891
от: 21889
во: 21300
ка: 19484
ол: 19141
ер: 18775
ло: 18707
ла: 18483
ли: 18198
ор: 18154
ро: 18131
ел: 17340
ан: 17235
пр: 16821
ре: 16024
од: 15276
```

```
Bigram frequency with crossing and without spaces:
то: 1.757246471408356 %
ст: 1.3560426859940717 %
на: 1.2482137293501891 %
ов: 1.1660148193016444 %
ал: 1.0696301788794491 %
го: 1.0612753928925138 %
не: 1.0553139049747529 %
ос: 1.049004300974203 %
но: 1.047046148008515 %
ен: 1.0289876262138373 %
он: 1.0284219375793051 %
по: 1.0127567138538016 %
ко: 1.0053157325841875 %
ра: 0.9729844513951622 %
ни: 0.9525761460416591 %
от: 0.9524891170209617 %
во: 0.9268590704256242 %
ка: 0.8478367196325286 %
ол: 0.8329112425829517 %
ер: 0.8169849317953565 %
ло: 0.8140259450916503 %
ла: 0.8042786947735591 %
ли: 0.7918770593242022 %
ор: 0.7899624208688629 %
ро: 0.7889615871308446 %
ел: 0.7545416094450855 %
ан: 0.7499725858584804 %
пр: 0.7319575785741513 %
ре: 0.6972765138263005 %
од: 0.6647276600855321 %
```

1.5 Кількість та частота біграм без перетинів, але з пробілами:



```
BiGram with no crossing and with spaces:
о : 29644
и : 26048
е : 23013
а : 22764
 с: 21769
 п: 21220
 н: 21158
 в: 20292
то: 19707
 о: 16425
 и: 15365
ст: 15358
я : 14706
на: 14339
ь : 13522
 к: 13420
го: 12147
не: 12034
ал: 11978
но: 11684
по: 11571
ра: 11139
ко: 10988
ов: 10450
ни: 10429
 д: 10285
й : 10221
во: 10170
```

```
Bigram frequency with spaces and no crossing:
о : 2.16220947407194 %
и : 1.8999201315823062 %
е : 1.6785496770617174 %
а : 1.660387817695778 %
 с: 1.5878133194262603 %
 п: 1.5477697017881045 %
 н: 1.5432474717451796 %
 в: 1.4800821295327151 %
то: 1.437412700901893 %
 о: 1.198026265403846 %
 и: 1.1207107195086816 %
ст: 1.1202001451489965 %
я : 1.0726437905040462 %
на: 1.0458751062177016 %
ь : 0.9862837845230324 %
 к: 0.9788439867104789 %
го: 0.8859923924420408 %
не: 0.8777502634928392 %
ал: 0.8736656686153588 %
но: 0.8522215455085868 %
по: 0.8439794165593854 %
ра: 0.8124696846473937 %
ко: 0.8014558663170447 %
ов: 0.7622145798155368 %
ни: 0.7606828567364816 %
 д: 0.7501796127658178 %
й : 0.7455115043344116 %
во: 0.7417916054281348 %
```

1.6 Кількість та частота біграм з перетинами та з пробілами:

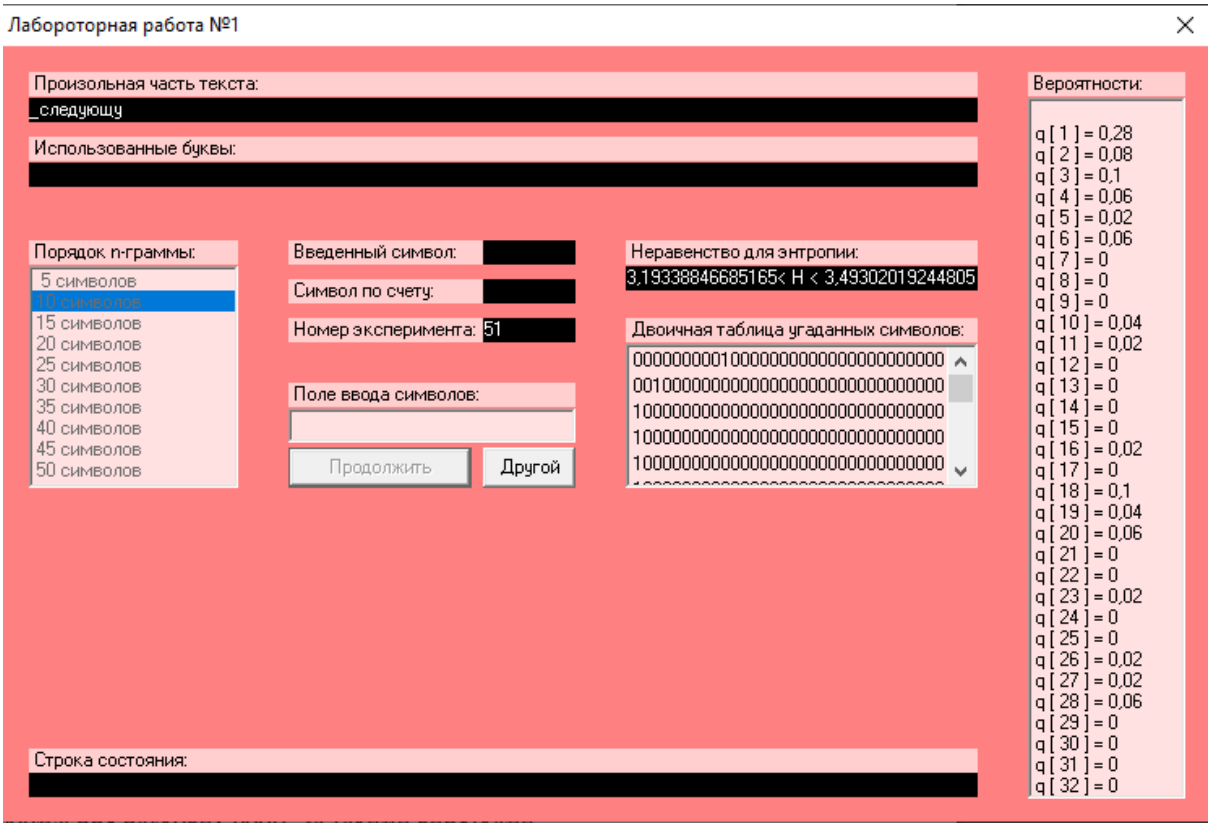
```
BiGram with crossing and spaces:
о : 59358
и : 52237
е : 46030
а : 45491
 с: 43324
 п: 42405
 н: 42402
 в: 40699
то: 39604
 о: 33019
 и: 30818
ст: 30406
я : 29593
на: 28578
 к: 26929
ь : 26912
го: 24215
не: 24072
ал: 23882
но: 23322
по: 23267
ра: 22261
ко: 22145
ни: 21069
```

```
Bigram frequency with spaces and crossing:
о : 2.1647623458703653 %
и : 1.9050623447762773 %
е : 1.6786955554501988 %
а : 1.6590384426023244 %
 с: 1.5800088256425031 %
 п: 1.546493265888892 %
 н: 1.5463838570975306 %
 в: 1.4842761332015566 %
то: 1.4443419243547617 %
 о: 1.204189627317187 %
 и: 1.1239200440552732 %
ст: 1.1088945700416848 %
я : 1.0792447875828317 %
на: 1.0422281465056655 %
 к: 0.9820897808541909 %
ь : 0.9814697977031448 %
го: 0.8831112942695322 %
не: 0.8778961418813206 %
ал: 0.8709669184284521 %
но: 0.8505439440410502 %
по: 0.8485381161994303 %
ра: 0.8118497014963475 %
ко: 0.8076192282303857 %
ни: 0.7683779417288777 %
```

1.7 H1 та H2:

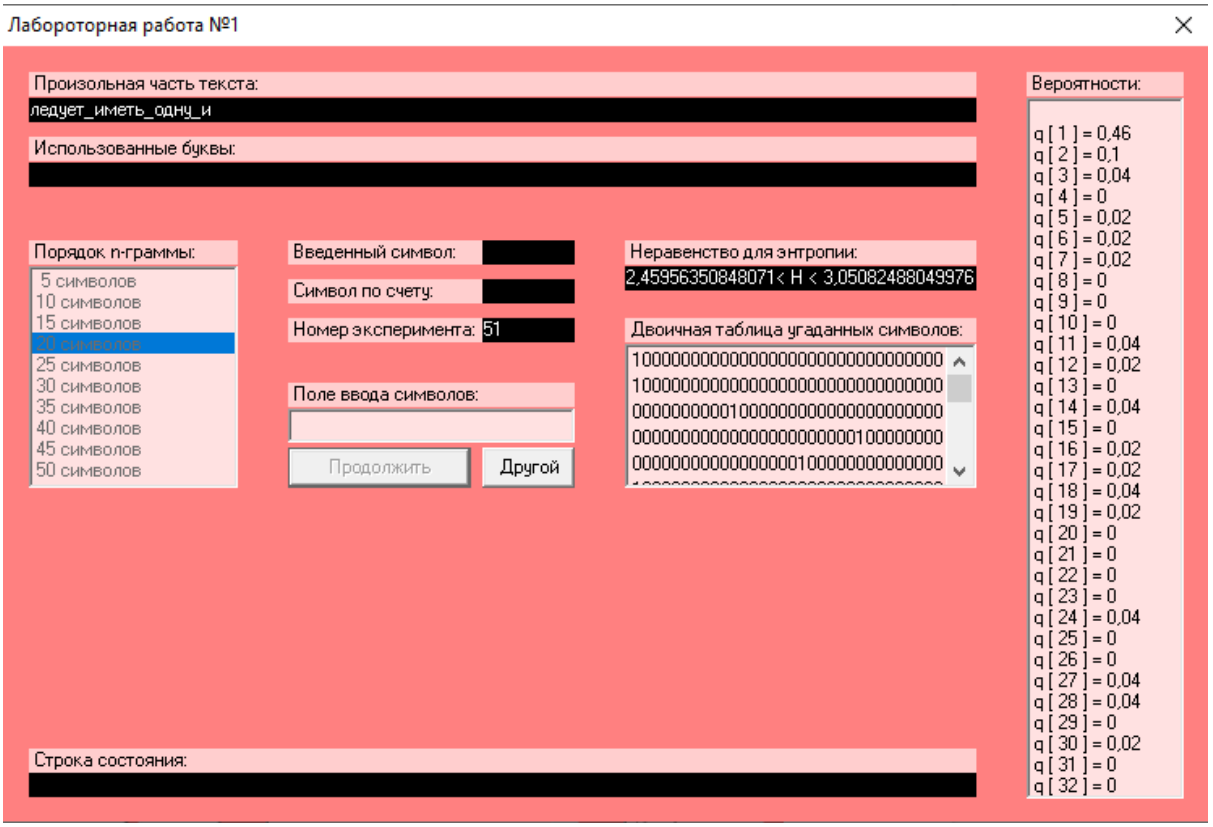
```
Entropy MonoGram without spaces: 4.459495560105348


Entropy MonoGram with spaces: 4.376343121574429


Entropy BiGram without spaces and crossing: 4.138908270465847


Entropy BiGram without spaces and with crossing: 4.138711256669181


Entropy BiGram with spaces and without crossing: 3.966535460585095


Entropy BiGram with spaces and crossing: 3.966562055656355
```

2. H(10)



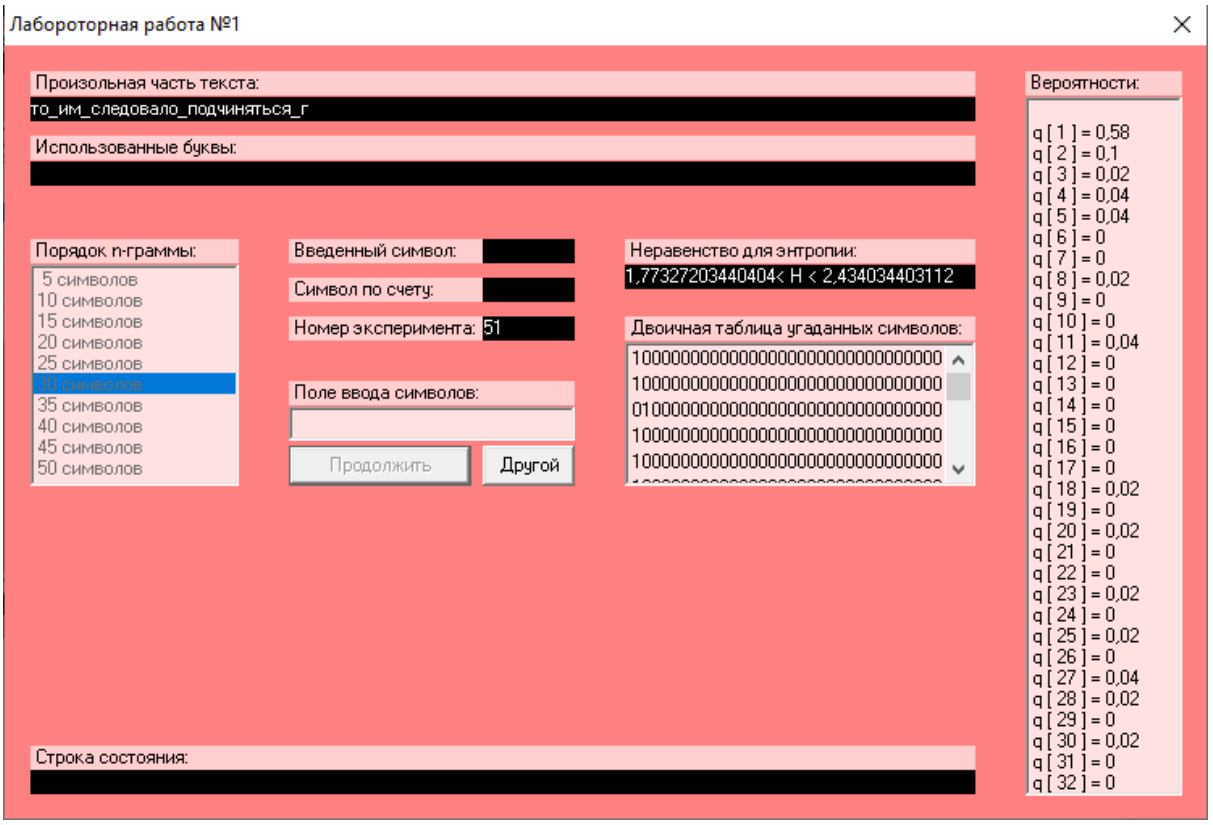
R = 0.33137

H(20)


Лабороторная работа №1
Произольная часть текста:
ледует_иметь_одну_и
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2,45956350848071< H < 3,05082488049976
Двоичная таблица угаданных символов:
Вероятности:
q[1] = 0,46
q[2] = 0,1
q[3] = 0,04
q[4] = 0
q[5] = 0,02
q[6] = 0,02
q[7] = 0,02
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0,04
q[12] = 0,02
q[13] = 0
q[14] = 0,04
q[15] = 0
q[16] = 0,02
q[17] = 0,02
q[18] = 0,04
q[19] = 0,02
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0,04
q[25] = 0
q[26] = 0
q[27] = 0,04
q[28] = 0,04
q[29] = 0
q[30] = 0,02
q[31] = 0
q[32] = 0
Строка состояния:

R = 0.44897

H(30)



R = 0.57928

Висновок

Виконую цю лабораторну роботу, ми дізналися ~~про існування російської мови.~~ Також виявилось, що у творі “Війна і мир” майже половина символів - це розділові знаки, латиниця та цифри. Тим самим з трохи більше 5 мільйонів символів оригінального тексту, після обробки, у нас залишилось приблизно 2.3 мільйони символів кирилиці.