НТУУ "КПІ ім Ігоря Сікорського"
Фізико-технічний інститут

КРИПТОГРАФІЯ
КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
студенти групи ФБ-14
Сергеєв Олег
Деркач Семен
Перевірила:
Селюх П.В.

Експериментальна оцінка ентропії на символ джерела
відкритого тексту
Варіант № 3

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку 1H та 2H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1H та2H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1H та2H на тому ж тексті, в якому вилучено всі пробіли.

Частоти для монограм, без пробілів:

```
 1 а ~ 0.07822784712620215
 2 б ~ 0.02208340921708328
 3 в ~ 0.04577069370494087
 4 г ~ 0.017297093271812017
 5 д ~ 0.030778999192835325
 6 е ~ 0.08640953675790507
 7 ж ~ 0.011073066835105666
 8 з ~ 0.01614604453546215
 9 и ~ 0.06732754764444919
10 й ~ 0.009965485044344538
11 к ~ 0.04191424026656792
12 л ~ 0.04452830936140454
13 м ~ 0.029992742670730183
14 н ~ 0.06301276552664056
15 о ~ 0.11294748257608228
16 п ~ 0.029731390782696633
17 р ~ 0.04145701201613449
18 с ~ 0.050144348774369685
19 т ~ 0.06188097429783001
20 у ~ 0.02808624942433808
21 ф ~ 0.001060813558165651
22 х ~ 0.009348144374084237
23 ц ~ 0.0034393908465215166
24 ч ~ 0.01933343714046921
25 ш ~ 0.009278817346942705
26 щ ~ 0.003066345414759944
27 ы ~ 0.018758462986795402
28 ь ~ 0.02029411165943043
29 э ~ 0.0025320871341692557
30 ю ~ 0.0057200299536774414
31 я ~ 0.018393120558049557
```

```
*** Ентропія монограми без пробілів ***
4.463308520322652
```

```
*** Надлишковість монограми без пробілів ***
0.09908525203876895
```

Частоти для монограм, з пробілами:

```
 1    ~ 0.16341132557661034
 2 а  ~ 0.06544453093030503
 3 б  ~ 0.018474730043668966
 4 в  ~ 0.03829124397405547
 5 г  ~ 0.014470552331642947
 6 д  ~ 0.025749362134812687
 7 е  ~ 0.07228923981383496
 8 ж  ~ 0.009263602305382647
 9 з  ~ 0.013507597995103295
10 и  ~ 0.056325463836047356
11 й  ~ 0.008337011923234312
12 к  ~ 0.03506497870407152
13 л  ~ 0.03725187930297204
14 м  ~ 0.025091588833228002
15 н  ~ 0.05271576598368409
16 о  ~ 0.09449058472778359
17 п  ~ 0.024872944803659962
18 р  ~ 0.03468246672813249
19 с  ~ 0.04195019427097406
20 т  ~ 0.051768922259849456
21 у  ~ 0.023496638175431683
22 ф  ~ 0.0008874646084361613
23 х  ~ 0.0078205517102336
24 ц  ~ 0.0028773554291153755
25 ч  ~ 0.016174134549393065
26 ш  ~ 0.007762553504495551
27 щ  ~ 0.002565269845858261
28 ы  ~ 0.015693117684343384
29 ь  ~ 0.01697782397176316
30 э  ~ 0.002118315419099177
31 ю  ~ 0.004785312276609093
32 я  ~ 0.015387476346168275
```

```
*** Ентропія монограми з пробілами ***
4.376361783687937
```

```
*** Надлишковість монограми з пробілами ***
0.12472764326241259
```

Частоти для біграм, без пробілів(перетинаються):

```
 1 ир ~ 0.0011521492
 2 яу ~ 0.0002503476
 3 зо ~ 0.0008330248
 4 ез ~ 0.0020616537
 5 кф ~ 1.59562e-05
 6 сб ~ 0.000174418
 7 же ~ 0.0048396868
 8 бг ~ 3.02618e-05
 9 йв ~ 0.0005315072
10 зы ~ 0.0004506257
11 иу ~ 0.0005337081
12 ьу ~ 0.0002850111
13 юф ~ 3.35631e-05
14 бы ~ 0.0050713271
15 вз ~ 0.0006343973
16 ны ~ 0.0040600328
17 чз ~ 6.6026e-06
18 хи ~ 0.0005529656
19 лк ~ 0.000810466
20 йй ~ 5.502e-07
21 зт ~ 0.0001094927
22 зк ~ 0.0002630025
23 ьд ~ 0.0006222926
24 яф ~ 2.75107e-05
25 ыг ~ 0.0002459459
26 эф ~ 3.3013e-06
27 ыш ~ 0.0005166514
28 хт ~ 0.0002332909
29 ьз ~ 0.0005122497
30 ущ ~ 0.0003136223
31 ту ~ 0.001828913
32 шт ~ 7.37287e-05
33 цк ~ 8.03313e-05
34 яд ~ 0.0011284899
35 жс ~ 9.62875e-05
36 тц ~ 5.8873e-05
37 ям ~ 0.0007570952
38 ню ~ 0.0001524094
39 ря ~ 0.0011356427
40 уч ~ 0.001510889
41 чр ~ 7.04275e-05
42 хя ~ 1.92575e-05
43 от ~ 0.0087951788
44 ыч ~ 0.0002685047
45 зу ~ 0.0003917527
46 уд ~ 0.0026784442
47 цп ~ 0.0001111433
48 ьв ~ 0.0012726461
49 сь ~ 0.0032072003
50 сш ~ 0.0001111433
51 пэ ~ 0.76069e-05
52 ек ~ 0.0030476381
```

```
*** Ентропія біграми без пробілів з перетином ***
4.134798718566218
```

```
*** Надлишковість біграми без пробілів з перетином ***
0.16539465545657184
```

Всього 874 біграми*

Частоти для біграм, без пробілів(не перетинаються):

```
 1 ир ~ 0.001160952
 2 яу ~ 0.0002630024
 3 зо ~ 0.0008341247
 4 ез ~ 0.0020280895
 5 кф ~ 1.65064e-05
 6 сб ~ 0.0001760685
 7 же ~ 0.0048440858
 8 бг ~ 2.53099e-05
 9 йв ~ 0.0005392099
10 зы ~ 0.0004610795
11 иу ~ 0.0005436116
12 ьу ~ 0.0002883122
13 юф ~ 2.53099e-05
14 бы ~ 0.0051136908
15 вз ~ 0.0006569558
16 ны ~ 0.0040506769
17 чз ~ 7.703e-06
18 хи ~ 0.0005271052
19 лк ~ 0.0008198192
20 зт ~ 0.0001023398
21 зк ~ 0.0002663037
22 ьд ~ 0.0006338468
23 яф ~ 2.97116e-05
24 ыг ~ 0.0002420942
25 эф ~ 5.5021e-06
26 ыш ~ 0.0005194022
27 хт ~ 0.0002431947
28 ьз ~ 0.0005293061
29 ущ ~ 0.0002751071
30 ту ~ 0.0017661876
31 шт ~ 7.703e-05
32 цк ~ 8.25321e-05
33 яд ~ 0.0011576507
34 жс ~ 0.0001111433
35 тц ~ 5.94231e-05
36 ям ~ 0.0007471909
37 ню ~ 0.0001408548
38 ря ~ 0.0011158344
39 уч ~ 0.001505386
40 чр ~ 7.15278e-05
41 хя ~ 2.09081e-05
42 от ~ 0.0088078289
43 ыч ~ 0.0002475964
44 зу ~ 0.0003884512
45 уд ~ 0.0026058144
46 цп ~ 0.0001221476
47 ьв ~ 0.0012963047
48 сь ~ 0.0032649711
49 сш ~ 0.000115545
50 пэ ~ 1.76069e-05
51 ек ~ 0.0030272785
```

```
*** Ентропія біграми без пробілів без перетину ***
4.1340462834770335
```

```
*** Надлишковість біграми без пробілів без перетину ***
0.1655465337920361
```

Частоти для біграм, з пробілами(перетинаються):

```
 1 ир ~ 0.0005265869
 2 яу ~ 1.3809e-06
 3 зо ~ 0.0006430436
 4 ез ~ 0.0010826332
 5 кф ~ 3.2221e-06
 6 сб ~ 5.1554e-05
 7  ж ~ 0.0019585902
 8 же ~ 0.0040359386
 9 бг ~ 5.5236e-06
10 зы ~ 0.0003769883
11 йв ~ 5.5236e-06
12 ьу ~ 3.2221e-06
13 иу ~ 1.19679e-05
14 юф ~ 1.01267e-05
15 бы ~ 0.0042426148
16 вз ~ 0.0003852738
17 ны ~ 0.0033965775
18 хи ~ 0.000138091
19 лк ~ 0.000352132
20 зт ~ 1.84121e-05
21 зк ~ 0.0001054094
22 ьд ~ 4.74112e-05
23 яф ~ 4.603e-07
24 ыг ~ 0.0001242819
25 эф ~ 2.7618e-06
26 ш  ~ 0.0002458019
27 ыш ~ 0.0004230187
28 хт ~ 3.36021e-05
29 ьз ~ 0.0002117395
30 ущ ~ 0.0002586904
31 ту ~ 0.0013657197
32 л  ~ 0.0059848625
33 шт ~ 5.06334e-05
34 цк ~ 2.39358e-05
35 яд ~ 0.0004340659
36 жс ~ 4.23479e-05
37 тц ~ 3.82052e-05
38 ям ~ 0.0004004638
39 ню ~ 0.000127504
40 ря ~ 0.000942701
41 уч ~ 0.0008589258
42 з  ~ 0.0011710114
43 чр ~ 5.70776e-05
44 от ~ 0.0062951068
45 ыч ~ 0.0001095522
46 зу ~ 0.0002890704
47 уд ~ 0.0019816054
48 ьв ~ 1.88724e-05
49 цп ~ 9.6664e-06
50  у ~ 0.0038821974
51  о ~ 0.0091940965
52 сь ~ 0.0026831075
53 сш ~ 8.33149e-05
```

```
*** Ентропія біграми з пробілами з перетином ***
3.972166407711179
```

```
*** Надлишковість біграми з пробілами з перетином ***
0.2055667184577642
```

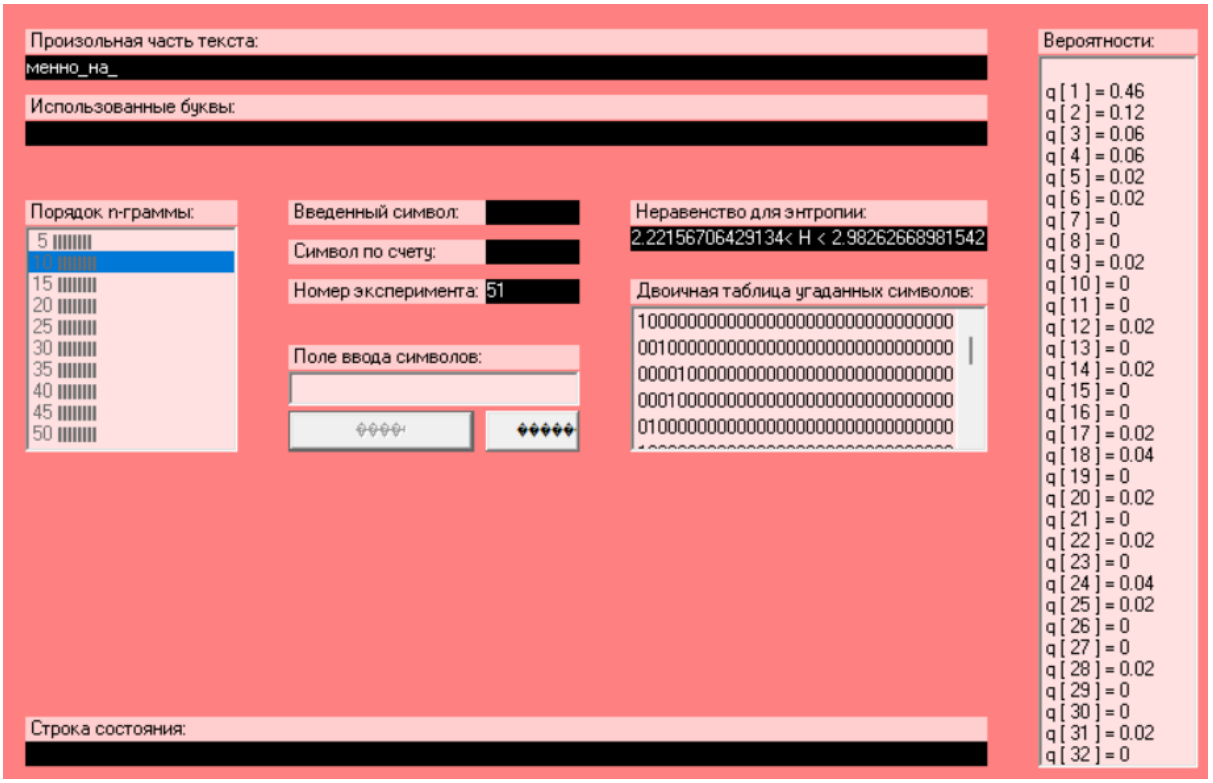
Частоти для біграм, без пробілів(не перетинаються):

```
 1 ир ~ 0.0005459194
 2 яу ~ 1.8412e-06
 3 ез ~ 0.0010946006
 4 зо ~ 0.0006241709
 5 кф ~ 2.7618e-06
 6 сб ~ 5.33951e-05
 7  ж ~ 0.0018872423
 8 же ~ 0.0040948556
 9 бг ~ 3.6824e-06
10 зы ~ 0.0003802103
11 йв ~ 5.5236e-06
12 ьу ~ 4.603e-06
13 иу ~ 5.5236e-06
14 юф ~ 1.01267e-05
15 бы ~ 0.0042347877
16 вз ~ 0.0003857339
17 ны ~ 0.0034264956
18 хи ~ 0.0001390115
19 лк ~ 0.000369163
20 зт ~ 2.11739e-05
21 зк ~ 9.39018e-05
22 ьд ~ 5.43158e-05
23 яф ~ 9.206e-07
24 ыг ~ 0.0001288848
25 эф ~ 9.206e-07
26 ш  ~ 0.0002439606
27 ыш ~ 0.0004188757
28 хт ~ 3.4983e-05
29 ьз ~ 0.000221866
30 ущ ~ 0.0002596109
31 ту ~ 0.0013744648
32 л  ~ 0.0059461943
33 шт ~ 4.78715e-05
34 цк ~ 2.5777e-05
35 яд ~ 0.0004474145
36 жс ~ 4.23479e-05
37 тц ~ 4.14273e-05
38 ям ~ 0.0003774485
39 ню ~ 0.0001344085
40 ря ~ 0.0009288915
41 уч ~ 0.0008745757
42 з  ~ 0.0011774551
43 чр ~ 5.6157e-05
44 от ~ 0.0063503403
45 ыч ~ 0.0001086315
46 зу ~ 0.0002881497
47 уд ~ 0.0019618114
48 ьв ~ 1.74915e-05
49 цп ~ 1.10473e-05
50  у ~ 0.0038232768
51  о ~ 0.0092235517
52 сь ~ 0.0026642338
53 сш ~ 8.00927e-05
```

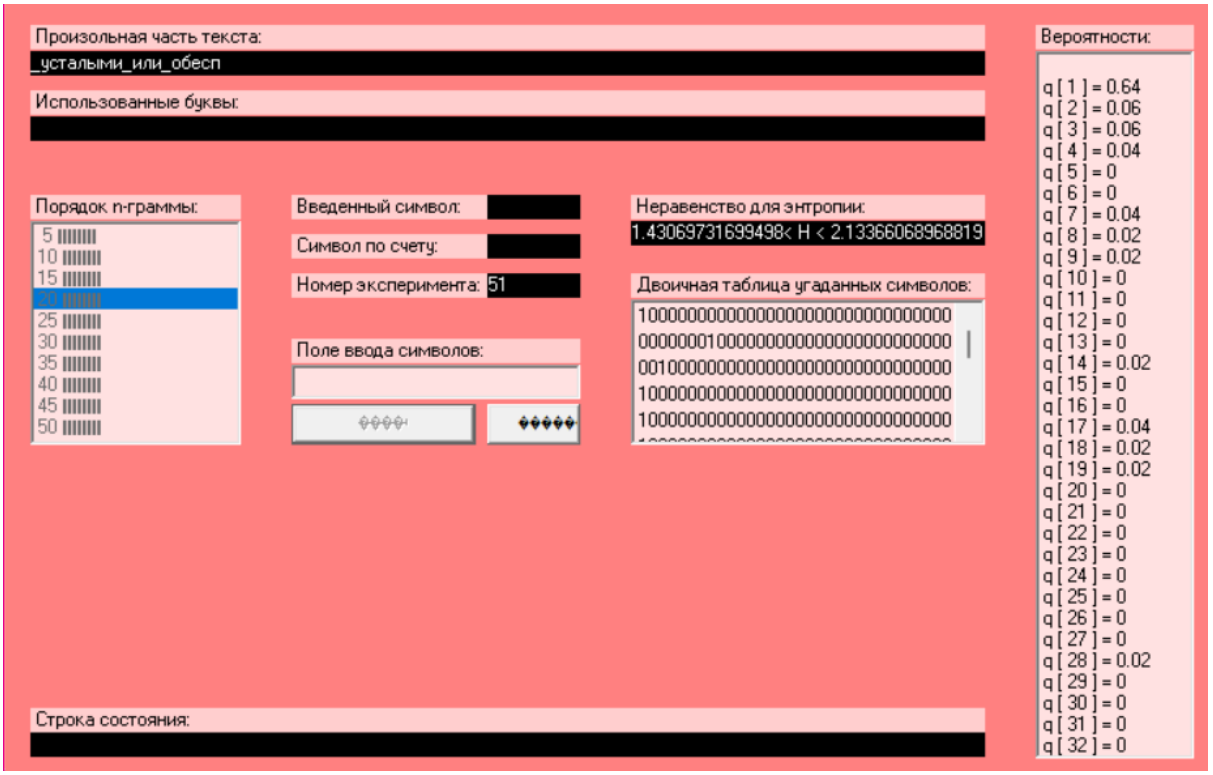
```
*** Ентропія біграми з пробілами без перетину ***
3.9716821966657596
```

```
*** Надлишковість біграми з пробілами без перетину ***
0.2056635606668481
```

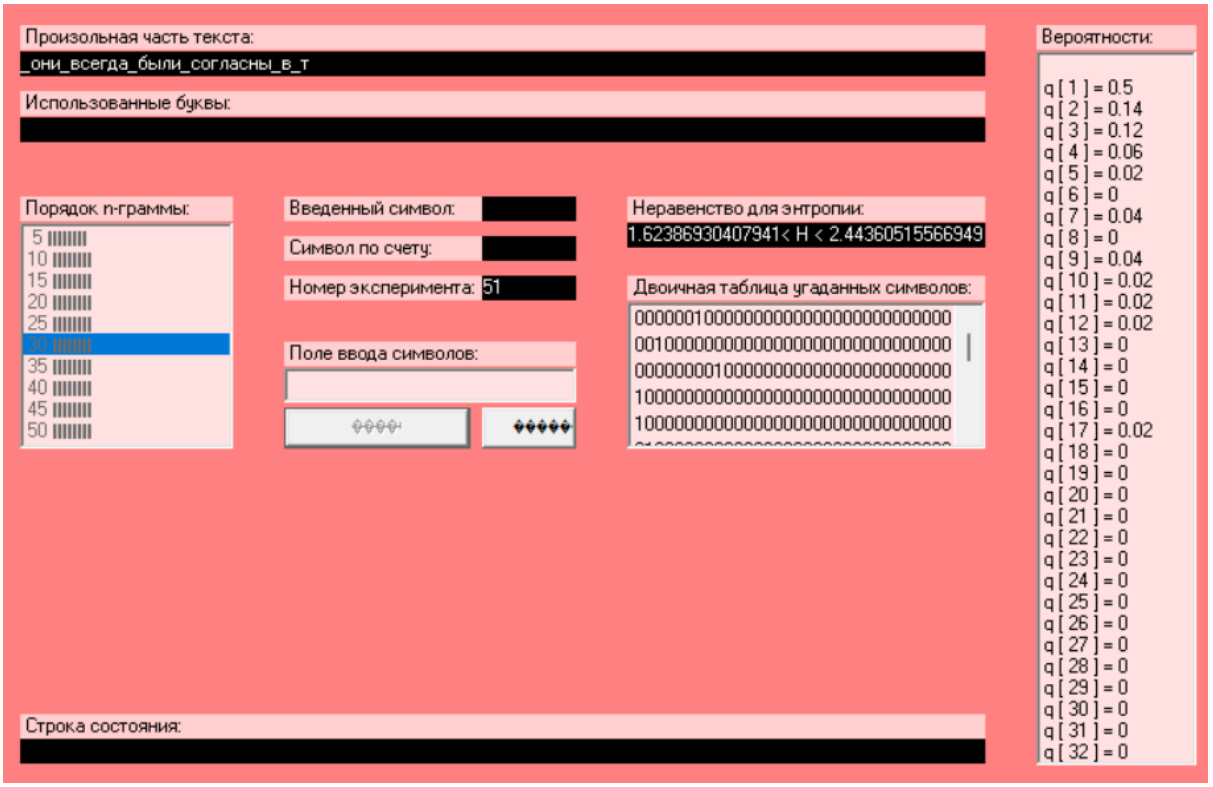
2. За допомогою програми CoolPinkProgram оцінити значення H (10) , H (20) , H (30) .
H(10):



H(20):



H(30):



3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

R(10):

$$1 - \frac{2.22156706429134}{5} = 0.555686587141732$$

$$1 - \frac{2.98262668981542}{5} = 0.403474662036916$$

$$0.403474662036916 < R(10) < 0.555686587141732$$

R(20):

$$1 - \frac{1.43069731699498}{5} = 0.713860536601004$$

$$1 - \frac{2.13366068968819}{5} = 0.573267862062362$$

$$0.555686587141732 < R(20) < 0.713860536601004$$

R(30):

$$1 - \frac{1.62386930407941}{5} = 0.675226139184118$$

$$1 - \frac{2.44360515566949}{5} = 0.511278968866102$$

$$0.511278968866102 < R(30) < 0.675226139184118$$

Висновок:

В ході виконання першого комп’ютерного практикуму ми визначили означення ентропії та надлишковості тексту, підвищили навички з процедурного програмування.

Найскладнішим елементом комп’ютерного практикуму була реалізація обробки тексту мовою Python, якщо конкретніше, то з визначенням ентропії для біграм: аналіз проблеми було розтягнуто на довгі години, хоча помилка закралась лише в формулі для обчислення ентропії. Крім цього варто відмітити специфіку CoolPinkProgram, ентропія для 10-грамми постійно була доволі високою, адже не завжди виходило відгадати символ з короткого набору слів, але в результаті значення можна назвати “нормальним”.