КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела

відкритого тексту

Виконав

ФБ-12 Сущенко Олександр

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 1 H та 2 H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 H та 2 H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 H та 2 H 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

Створюємо програму, яка спочатку буде фільтрувати за критеріями текст, створюючи два нових з пробілами та без відповідно. Далі обчислюємо значення частот символів та їх ентропію, використовуючи формулу:

$$H_1 = -\sum_{i=1}^{n} p(i) \log_2 p(i)$$

Далі рахуємо частоти біграм з перетинами та без; з пробілами та без. Наступний етап – обчислення ентропії, використовуємо наступну формулу:

$$H_2 = -\sum i,j\, p(i,j) \log_2 p(i,j)/2$$

Потім рахуємо надлишковість за допомгою іншої формули:

$$R = 1 - \frac{H_\infty}{H_0}$$

Результати виконання:

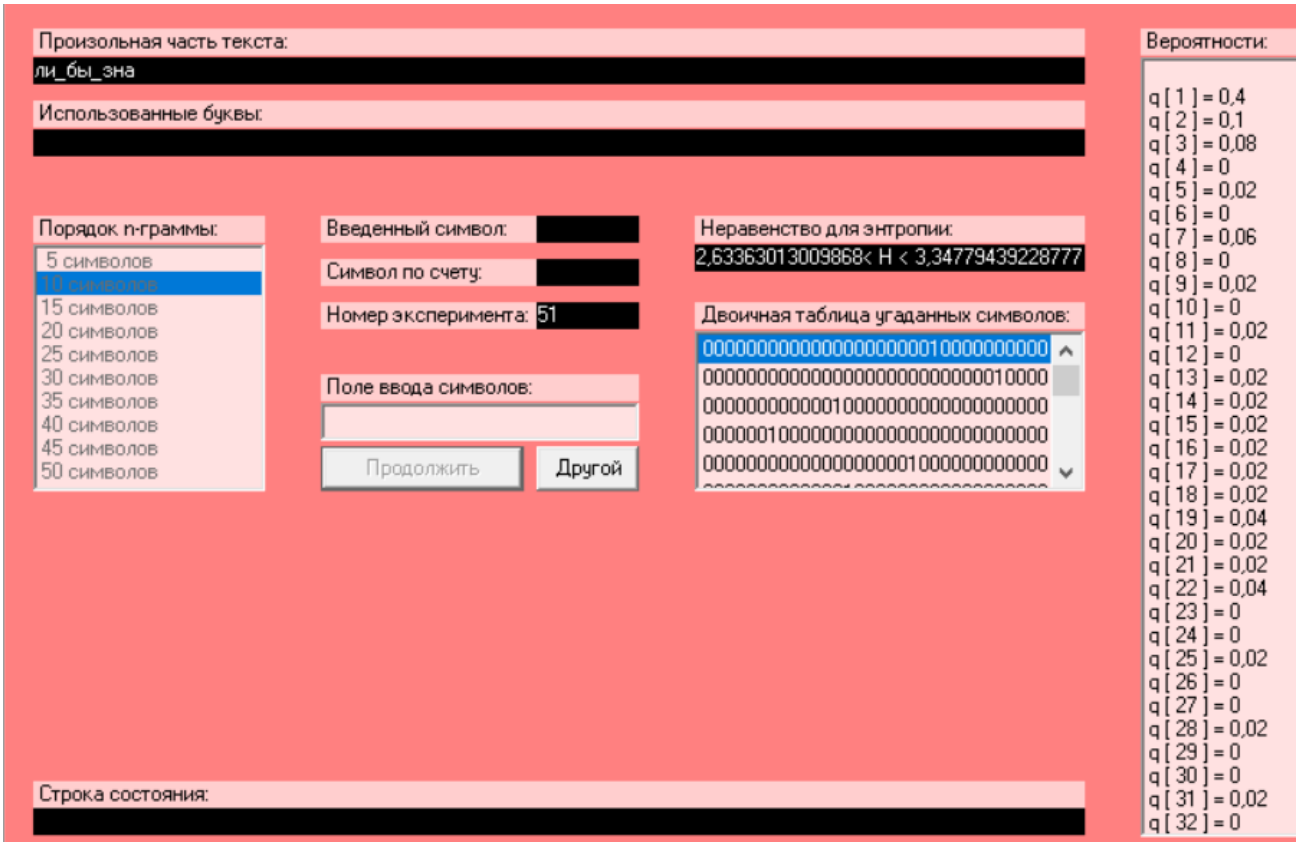
```
Частота символів без пробілів:
о 0.11223137474889013
е 0.08425977281232766
а 0.0795172677347457
и 0.07499155384689636
н 0.06438302730698661
т 0.05865942341455618
с 0.05533713799233952
л 0.04760544636435636
р 0.04738865517462372
в 0.04736808374786077
м 0.032534502634329436
к 0.03093626101659244
д 0.03041801930391039
п 0.02819709642223328
у 0.02523639338581157
я 0.019904437810598873
ы 0.019145668646534626
г 0.01809336104673749
ь 0.017603602848034916
б 0.016868569945620224
з 0.016729317210609475
ч 0.012958416443216136
й 0.011432966027871118
х 0.010306284808238699
ж 0.009619515637844779
ш 0.007282285074084829
ю 0.0060923071567202496
ц 0.005410285238656243
щ 0.0037985930726511584
э 0.0035588568299906084
ф 0.0018632965394904141
ъ 0.00026584613047506546
ё 2.3736261649559415e-06
```

```
Частота символів з пробілами:
  0.15449145620195515
о 0.09489258623238676
е 0.07124235781130525
а 0.06723252924920409
и 0.06340599949024202
н 0.05443639966364001
т 0.049597043671274335
с 0.04678802296185445
л 0.04025081163238288
р 0.040067512829243784
в 0.040050119512157596
м 0.027508199945545538
к 0.026156873002695296
д 0.025718695206870094
п 0.023840885935295524
у 0.021337586222352153
я 0.0168293722283582
ы 0.016187826417371376
г 0.015298091351039284
ь 0.014883996609641115
б 0.014262520010676821
з 0.014144780633477988
ч 0.010956451816832316
й 0.009666670457517826
х 0.008714051860181813
ж 0.008133382658996659
ш 0.006157234248511701
ю 0.005151097752448945 5
ц 0.004574442393668297
щ 0.003211742897338622
э 0.003090438559110854
ф 0.001575433143768476 3
ъ 0.000224775174652317 62
ё 2.0069212022528358e-06
```

```
Частоти біграм без пробілу без перетину:
 {'ст': 9032, 'то': 8801, 'на': 7671, 'но': 7281, 'ен': 7012, 'ов': 6964, 'по': 6758, 'ни': 6708, 'ра': 6584, 'ос': 6423, 'ко': 6102, 'не': 6080, 'го': 5892, 'во': 5886, 'ли': 58
 {'ст': 0.014292405656197384, 'то': 0.013926866937576747, 'на': 0.012138733811856747, 'но': 0.011521590520679047, 'ен': 0.011095919891635967, 'ов': 0.01101996379426025, 'по': 0.0
Частоти біграм без пробілу з перетином:
 {'ст': 18079, 'то': 17823, 'на': 15387, 'но': 14631, 'ен': 14198, 'ов': 14034, 'по': 13551, 'ни': 13066, 'ра': 13060, 'ос': 12884, 'ко': 12155, 'не': 12103, 'во': 11788, 'го': 1
 {'ст': 0.01430427379641234, 'то': 0.014101724203410429, 'на': 0.012174338232501613, 'но': 0.011576183965667845, 'ен': 0.011233590318129455, 'ов': 0.011103831985112604, 'по': 0.0
Частоти біграм з пробілами без перетину:
 {'о ': 13916, 'и ': 13855, ' п': 12535, 'е ': 11747, ' с': 11461, ' в': 11177, 'а ': 11099, ' н': 10449, 'ст': 8894, 'то': 8684, ' и': 8282, ' о': 7786, 'я ': 7724, 'на': 7575,
 {'о ': 0.018618889422581626, 'и ': 0.018537274572425152, ' п': 0.016771182732973604, 'е ': 0.01571687942275556, ' с': 0.015334226190874389, ' в': 0.014954248855719662, 'а ': 0.0
Частоти біграм з пробілами з перетином:
 {'о ': 27892, 'и ': 27376, ' п': 24941, 'е ': 23721, ' с': 23009, ' в': 22539, 'а ': 21932, ' н': 20850, 'ст': 17819, 'то': 17326, ' и': 16852, ' о': 15613, 'я ': 15313, 'на': 1
 {'о ': 0.01865902787347825, 'и ': 0.018313837195767266, ' п': 0.016684885063545856, 'е ': 0.015868736561981126, ' с': 0.015392426944674497, ' в': 0.015078009079317592, 'а ': 0.0
```

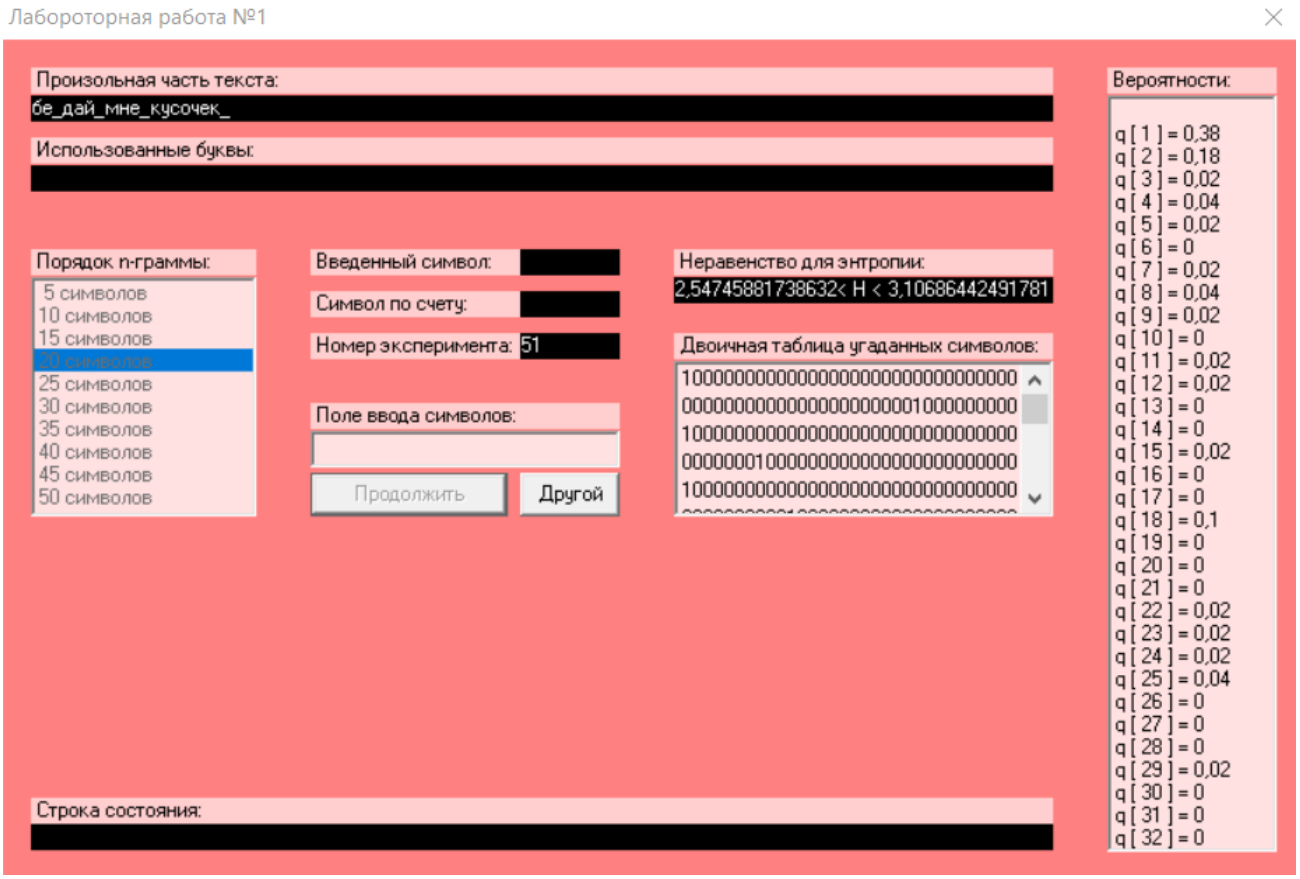
```
Ентропія H1 без пробілів:  4.457223768425505
Ентропія H1 з пробілами:  4.389587744688843
Ентропія H2 без пробілів без перетину:  4.149596664532538
Ентропія H2 без пробілів з перетином:  4.150057053053292
Ентропія H2 з пробілами без перетину:  3.9976662350370633
Ентропія H2 з пробілами з перетином:  3.9983739015880295

Надлишковість H1 без пробілів:  0.11640057002675763
Надлишковість H1 з пробілами:  0.13717546807476644
Надлишковість H2 без пробілів без перетину:  0.1773845250100553
Надлишковість H2 без пробілів з перетином:  0.17729325765269577
Надлишковість H2 з пробілами без перетину:  0.21421219971907213
Надлишковість H2 з пробілами з перетином:  0.2140730996267376
```

$H^{(10)}$:



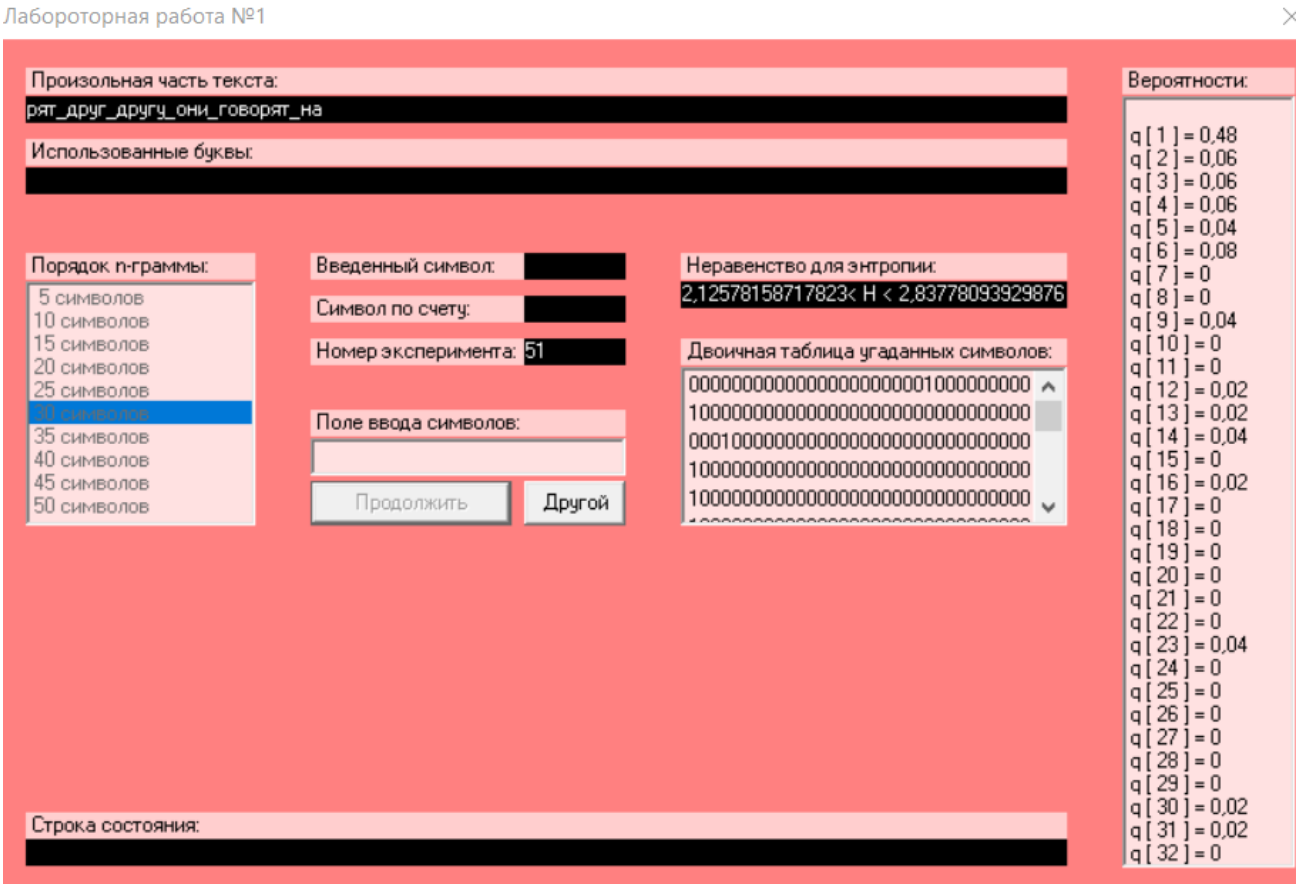
$0.330441121542446 < R < 0.473273973980264$

$H^{(20)}$:



$0.378627115016438 < R < 0.490508236522736$

$H^{(30)}$:



0,432443812140248 < R < 0.574843682564354

Висновки

Під час виконання цієї лабораторної роботи, я набув навичок оцінки частот букв та біграм. За допомогою практичної частини я навчився розраховувати ентропію та надлишковість. Також цікавою виявилася можливість застосування цих знань при роботі з CoolPinkProgram. Я використовував ці дані для передбачення наступного символу, вибираючи його на основі ймовірності, хоча в деяких випадках всеодно виникали складності.