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
студенти 3 курсу
групи ФБ-21
КАЮН Вероніка
РУДЮК 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.

Постановка задачі

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 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

Довільний текст

Спочатку підрахуємо частоту букв та біграм у нашому тексті **text2 –** текст з пробілами.

Частота букв	**Частота біграм *з перетином***	**Частота біграм *без перетину***

Буква	Частота
	0,123861
о	0,093987
и	0,088443
е	0,070679
н	0,063358
т	0,059321
а	0,057921
с	0,048501
р	0,040911
в	0,039834
л	0,034397
м	0,028153
к	0,025785
ы	0,022662
п	0,021747
д	0,021155
я	0,018087
у	0,017441
з	0,016956
г	0,01448
б	0,012704
ь	0,012327
х	0,012112
ч	0,009366
й	0,008236
ц	0,007913

Біграма	Частота
вв	5,38E-05
ве	0,002961
ед	0,002961
де	0,00323
ен	0,011789
ни	0,012865
ие	0,00576
е	0,015342
т	0,008559
те	0,009313
ех	0,004199
хн	0,00393
но	0,015288
ол	0,009366
ло	0,006298
ог	0,008398
ги	0,005975
ии	0,004952
и	0,024816
з	0,00183
за	0,003122
ан	0,008182
им	0,003337
ма	0,00253
аю	0,002153
ют	0,00366

Біграма	Частота
вв	5,57E-05
ед	0,002959
ен	0,011768
ие	0,005799
т	0,008594
ех	0,004236
но	0,015329
ло	0,006319
ги	0,00601
и	0,024799
за	0,003137
ни	0,012826
ма	0,002536
ют	0,003653
ц	0,001682
тр	0,004982
ал	0,004826
ьн	0,00317
ое	0,001188
м	0,005621
ес	0,00721
то	0,006916
в	0,009578
р	0,005494
аз	0,002625
ви	0,003014

Всі значення можна переглянути у файлі **frequency_data.xlsx**

Далі обчислюємо ентропію та надлишковість тексту.

```
Ентропія H1 (монограми): 4.61260
Надлишковість R1 (монограми): 0.23961
Ентропія H2 (біграми з перетином): 8.10212
Надлишковість R2 (біграми з перетином): 0.33218
Ентропія H2 (біграми без перетину): 8.10161
Надлишковість R2 (біграми без перетину): 0.33222
```

Текст без пробілів

Видаляємо пробіли із тексту та зберігаємо у **text3.**

Частота букв	**Частота біграм *з перетином***	**Частота біграм *без перетину***

Буква	Частота
о	0,104827
и	0,098643
е	0,078831
н	0,070665
т	0,066162
а	0,064601
с	0,054095
р	0,045629
в	0,044429
л	0,038365
м	0,0314
к	0,028758
ы	0,025276
п	0,024256
д	0,023595
	0,022813
я	0,020173
у	0,019452
з	0,018912
г	0,01615
б	0,014169
ь	0,013749
х	0,013509
ч	0,010447
й	0,009186
ц	0,008826

Біграма	Частота
вв	0,00024
ве	0,003302
ед	0,003842
де	0,003602
ен	0,014289
ни	0,014469
ие	0,006484
е	0,001081
т	0,002342
те	0,010387
ех	0,004743
хн	0,004503
но	0,017051
ол	0,010507
ло	0,007025
ог	0,009486
ги	0,006784
ии	0,008706
из	0,005403
за	0,003482
ан	0,009246
им	0,004983
ма	0,002942
аю	0,002402
ют	0,004203
тц	6E-05

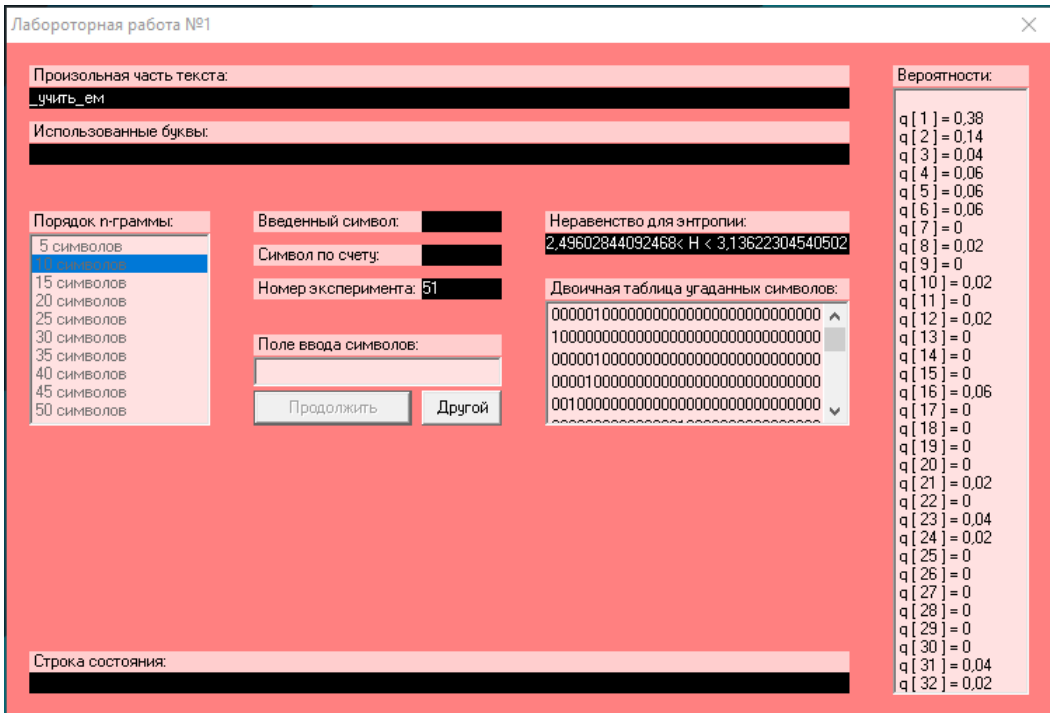
Біграма	Частота
вв	0,00012
ед	0,003722
ен	0,013329
ие	0,006004
т	0,002642
ех	0,005043
но	0,016691
ло	0,007085
ги	0,006724
из	0,005163
ан	0,008525
им	0,004323
аю	0,003002
тц	0,00012
тр	0,005283
ал	0,006364
ьн	0,004923
ое	0,001561
ме	0,004923
ст	0,020653
ов	0,013809
ра	0,012248
зв	0,002642
ит	0,005043
ии	0,008405
со	0,006004

Всі значення можна переглянути у файлі **frequency_data_spaces_del.xlsx**

```
Ентропія H1 (монограми): 4.63480
Надлишковість R1 (монограми): 0.23321
Ентропія H2 (біграми з перетином): 8.26551
Надлишковість R2 (біграми з перетином): 0.31627
Ентропія H2 (біграми без перетину): 8.26566
Надлишковість R2 (біграми без перетину): 0.31625
```

За допомогою програми CoolPinkProgram оцінимо значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$

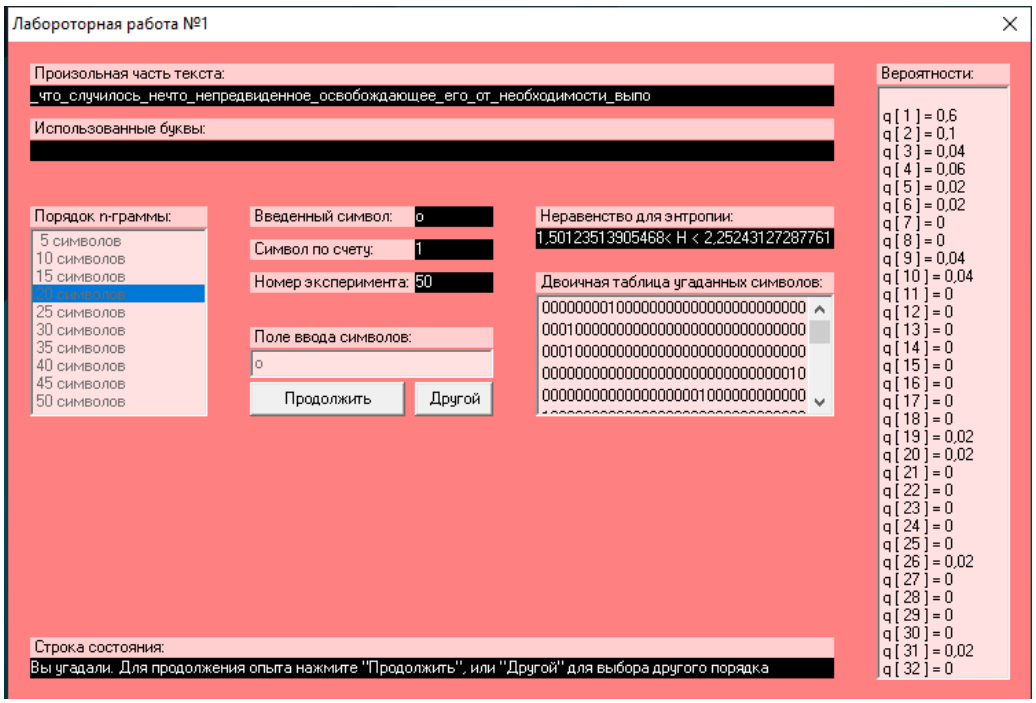
$$H^{(10)}$$



Надлишковість при мінімальному значенні ентропії складає: 0.5008

Надлишковість при максимальному значенні ентропії складає: 0.3728

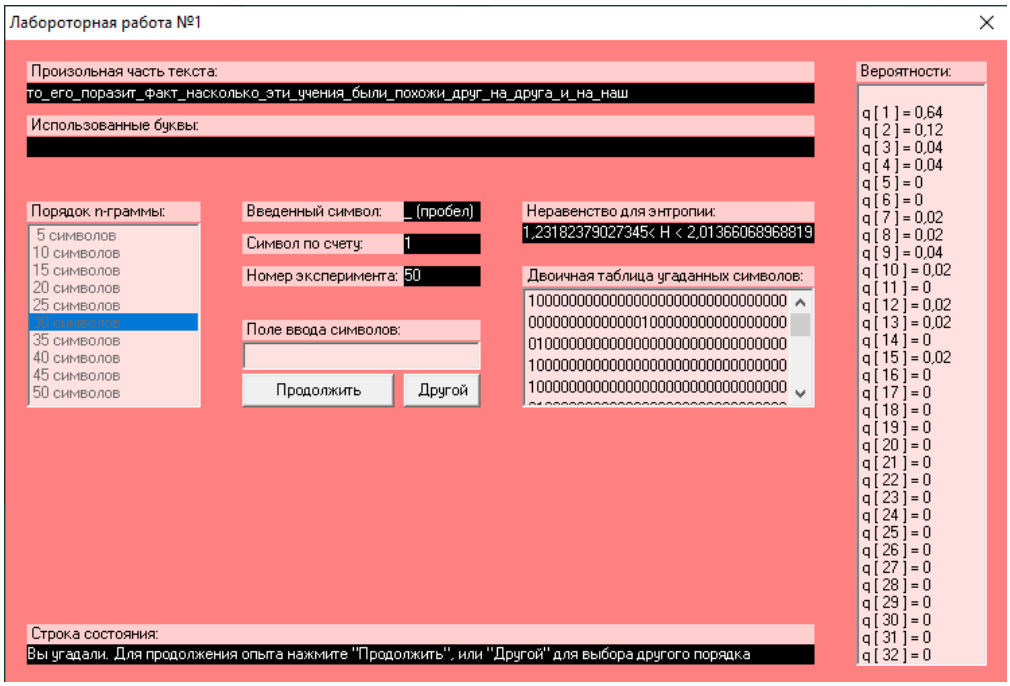
$$H^{(20)}$$



Надлишковість при мінімальному значенні ентропії складає: 0.6998

Надлишковість при максимальному значенні ентропії складає: 0.5495

$$H^{(30)}$$



Надлишковість при мінімальному значенні ентропії складає: 0.7536

Надлишковість при максимальному значенні ентропії складає: 0.5973

Висновок: під час виконання даної лабораторної роботи ми навчилися екпериментально визначати частоти літер і біграм у тексті і на основі цих значень обчислювати ентропію і надлишковіть. За допомогою програми CoolPinkProgram ми здійснили серію експериментів, щоб оцінити значення ентропії $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.