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

Виконав:
Дворніков Дмитро
Група:
ФБ-03

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

Хід роботи
Після написання програми, код якої буде разом з роботою ми отримали такі дані.

Частота букв з пробілами

```
 : 0.15872
о : 0.09504
е : 0.07071
а : 0.06699
н : 0.05671
и : 0.05656
т : 0.05604
с : 0.04454
л : 0.0406
р : 0.03849
в : 0.03524
м : 0.02777
к : 0.02723
д : 0.02405
у : 0.02376
п : 0.02351
я : 0.01723
ь : 0.0167
ы : 0.0163
г : 0.01496
з : 0.0145
б : 0.01353
ч : 0.01269
й : 0.00935
х : 0.00838
ж : 0.00824
ш : 0.00744
ю : 0.00445
щ : 0.00358
```

Тут ми бачимо, що пробіл є найчастішою "буквою" у російській мові, за ним ідуть всі нижче наведені.
Усі файли з даними будуть прикріплені разом з кодом та протоколом.

Частота букв без пробілів

```
о : 0.11297
е : 0.08405
а : 0.07963
н : 0.0674
и : 0.06723
т : 0.06661
с : 0.05294
л : 0.04826
р : 0.04575
в : 0.04189
м : 0.033
к : 0.03237
д : 0.02859
у : 0.02824
п : 0.02795
я : 0.02048
ь : 0.01985
ы : 0.01938
г : 0.01778
з : 0.01723
б : 0.01608
ч : 0.01509
й : 0.01111
х : 0.00996
ж : 0.0098
ш : 0.00884
ю : 0.00529
щ : 0.00426
```

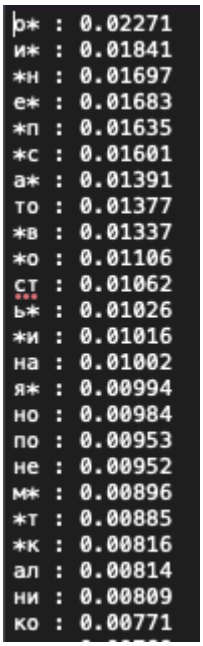
Нижче наведені значення про ентропію та надлишковість російської мови основане на частоті букв.

Ентропія та надлишковість російської мови

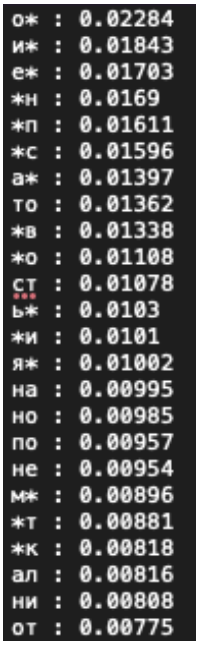
```
4.379441435294986 - Ентропія для тексту з пробілом

4.455388438097443 - Ентропія для тексту без пробілів

0.1391698432103622 - Надлишковість для тексту з пробілом

0.11676440565986546 - Надлишковість для тексту без пробілів
```

Далі буде наведені таблиці з частотами біграм з пробілами та без. Під час підрахунку значень були труднощі з написанням функції, так як я хотів зробити універсальну функцію. Але завдяки вкладці я зміг зрозуміти свою проблеми та вирішити її(проблема була у використанні мною неправильних методів для типу даних list)

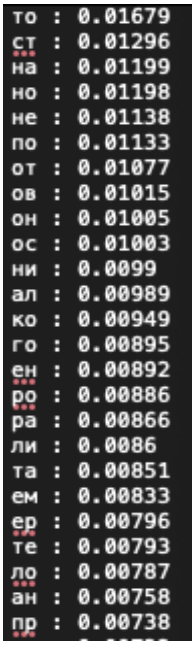
Частота перехресних біграм



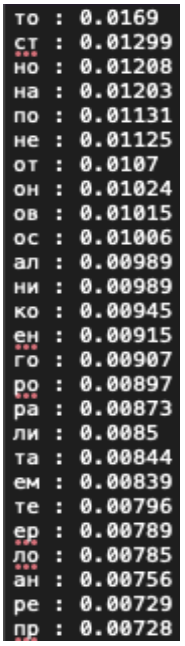
Частота не перехресних біграм



Частота перехресних біграм без пробілу



Частота не перехресних біграм без пробілу



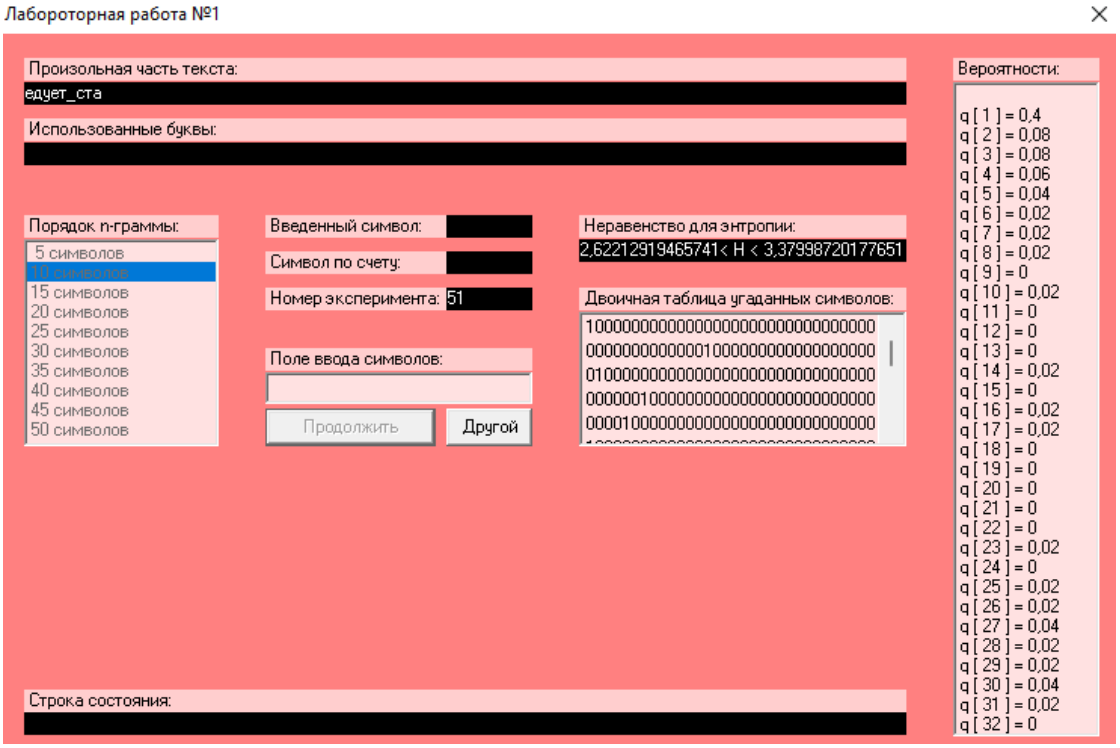
Також для всіх цих значень треба порахувати ентропію та надлишковість.

Ентропія та надлишковість для біграм

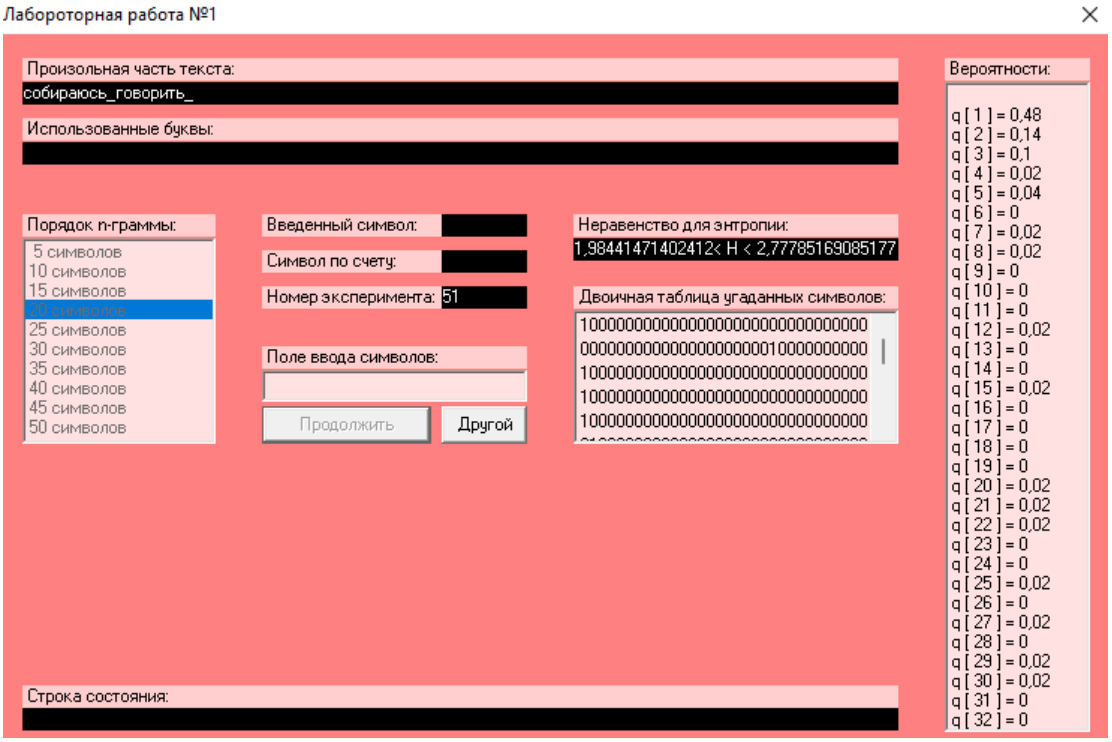
```
3.976565682256298 - Ентропія для перехресних біграм з пробілом

3.9765024430760287 - Ентропія для не перехресних біграм без пробілів

4.136699739436335 - Ентропія для перехресних біграм з пробілом

4.13635500411463 - Ентропія для не перехресних біграм без пробілів

0.21835975881467418 - Надлишковість для перехресних біграм з пробілом

0.21837218921116897 - Надлишковість для не перехресних біграм з пробілом

0.17994120967644744 - Надлишковість для перехресних біграм без пробілів

0.1800095499594523 - Надлишковість для не перехресних біграм без пробілів
```

CoolPinkProgram

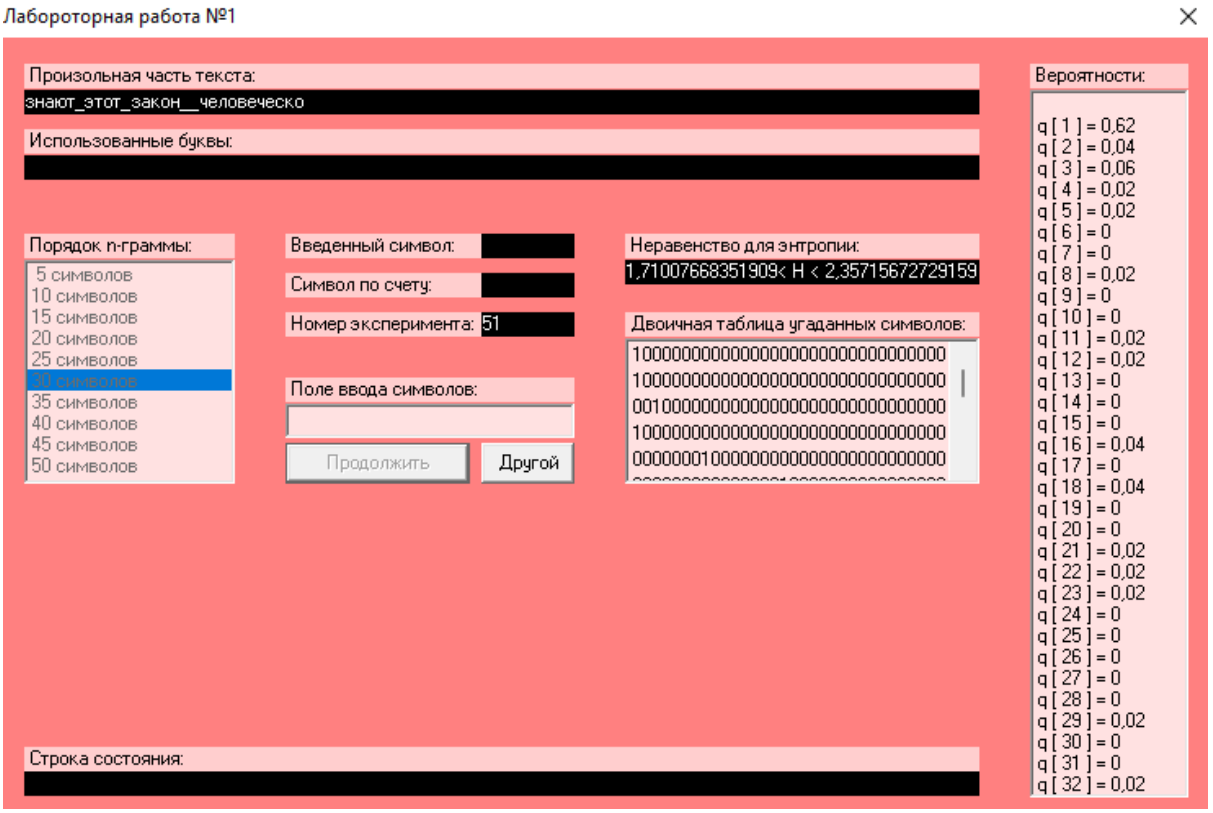
10 символів



20 символів



30 сиволів



Значення ентропії та надлишковості отриманні на прикладі роботи з CoolPinkProgram

Кіл-сть символів	H	R
10	3,001	0,90621875
20	2,38115	0,9255890625
30	2,0336	0,93645

Висновки

Під час роботи на цією лабораторною дізнався багато нового про мову, а саме про методи її аналізу такі як ентропія і надлишковість. Також практично порахував значення для вибраного тексту.