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”

Виконала
студентка групи ФБ-93
Куцовол Онисія
Перевірила
Селюх П.В.

Київ 2021

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.

Завдання:

- Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- 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.
- За допомогою програми CoolPinkProgram оцінити значення H(10), H(20) , H(30) .
- Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

Хід роботи:

Необхідно написати програму, що буде обчислювати частоти літер та біграм в тексті, а також рахувати ентропії за їх означенням. Найважчим під час виконання лабораторної роботи, було обрати, якою мовою програмувати. На момент отримання роботи, мені доводилося знайомитися лише з C++ на уроках програмування. Однак, трохи більше ознайомившись з мовою Python, та зрозумівши, що програмування нею буде набагато швидшим та простішим, вирішила використовувати її. Однак, на цьому труднощі не

закінчилися. Через те, що довелося зіштовхнутися з не зовсім знайомою мовою виникали деякі труднощі з підрахунком біграм, однак врешті-решт, вирішила для біграм, що не перетинаються використати цикл, що з кроком 2 ітерується по всьому тексту та додає одиницю до лічильника.

Нарешті отримуємо необхідні для нас результати.

Частоти літер з пробілами:

```
а  =>  0.07144139591174475
б  =>  0.015616895639105744
в  =>  0.036480429838677504
г  =>  0.015869585455706135
д  =>  0.025089438895612506
е  =>  0.06920043622242024
ж  =>  0.008913965767179583
з  =>  0.014117381069542898
и  =>  0.05361678924339349
й  =>  0.01180327432804458
к  =>  0.028211488077030496
л  =>  0.04120173956989533
м  =>  0.025418600630394594
н  =>  0.055152877339043234
о  =>  0.09317272013937838
п  =>  0.01990597278929659
р  =>  0.03467835246239576
с  =>  0.04465960021811121
т  =>  0.047921293771860995
у  =>  0.021797821547791623
ф  =>  0.0015128140335944461
х  =>  0.0
ц  =>  0.0
ч  =>  0.012554694045829953
ш  =>  0.008458459124097298
щ  =>  0.0023772791956484154
ы  =>  0.015763189743453337
ь  =>  0.01627854397467782
э  =>  0.0025169235679802105
ю  =>  0.006290646486946576
я  =>  0.018955061111037227
   =>  0.16733053157957734
```

Частоти літер без пробілів:

```
а  =>  0.0858287659193685
б  =>  0.01875968744507119
в  =>  0.04382719994886627
г  =>  0.019067288793723335
д  =>  0.03014493216791039
е  =>  0.08313625541298478
ж  =>  0.010710119684888384
з  =>  0.016962017225675526
и  =>  0.06442051101772103
й  =>  0.011489110113293172
к  =>  0.03389607069238267
л  =>  0.0495038430194468
м  =>  0.030540419616177435
н  =>  0.06626611910963391
о  =>  0.11195091161854237
п  =>  0.023917003563381858
р  =>  0.04166200604017194
с  =>  0.05365446381489589
т  =>  0.0575653952477589
у  =>  0.02619005768523993
ф  =>  0.001817644332944504
х  =>  0.0
ц  =>  0.0026925105063837266
ч  =>  0.015076460906664962
ш  =>  0.010162828973650149
щ  =>  0.0028562982374842204
ы  =>  0.018939454467010754
ь  =>  0.019554657164315047
э  =>  0.0030240807912944824
ю  =>  0.007558204566881322
я  =>  0.022770489445678402
```

Біграми з найвищими частотами:

- з пробілами і кроком 1

```
о  : 0.020345
и  : 0.019431
а  : 0.019078
 с : 0.017379
е  : 0.017063
 н : 0.016219
 п : 0.014131
 в : 0.014001
то : 0.013605
```

- з пробілами і кроком 2

```
о  : 0.020435
и  : 0.019583
а  : 0.018493
 с : 0.017502
е  : 0.017130
 н : 0.016126
 п : 0.014284
 в : 0.013811
то : 0.013599
```

- без пробілів з кроком 1:

```
то : 0.016702
на : 0.013331
ст : 0.013239
не : 0.011082
ал : 0.010898
он : 0.010670
ко : 0.010654
го : 0.010490
ла : 0.010395
```

- без пробілів з кроком 2:

```
то : 0.016555
на : 0.013455
ст : 0.013391
не : 0.011465
ал : 0.010938
он : 0.010834
го : 0.010762
ко : 0.010698
ов : 0.010163
```

Далі програма обчислює значення ентропії.

В тексті без пробілів:

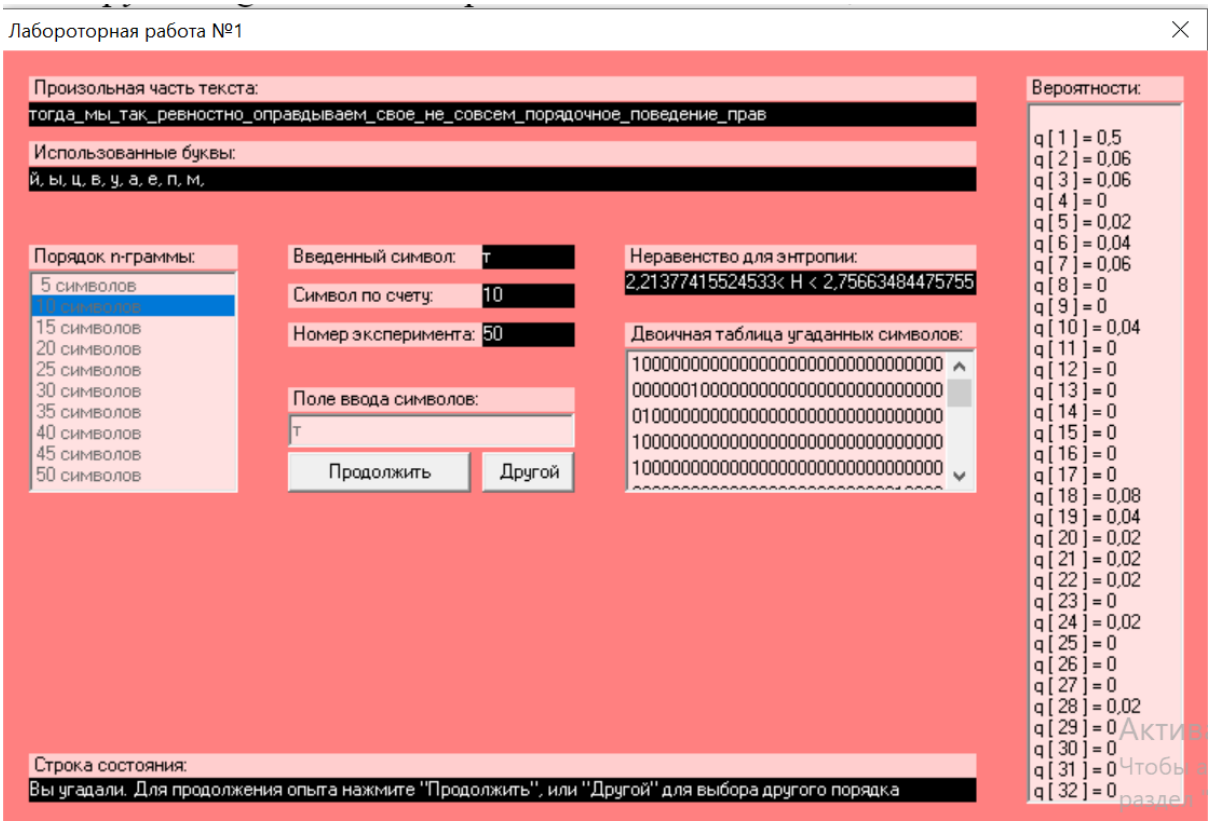
```
Letter Entropia:  4.540594180895466
Bigram 1-stepped:  4.180901854850425
Bigram 2-stepped:  4.179679885683603
```

В тексті з пробілами:

```
Letter Entropia:  4.424816926674272
Bigram 1-stepped:  3.9993347075470806
Bigram 2-stepped:  3.995740109609226 7
```

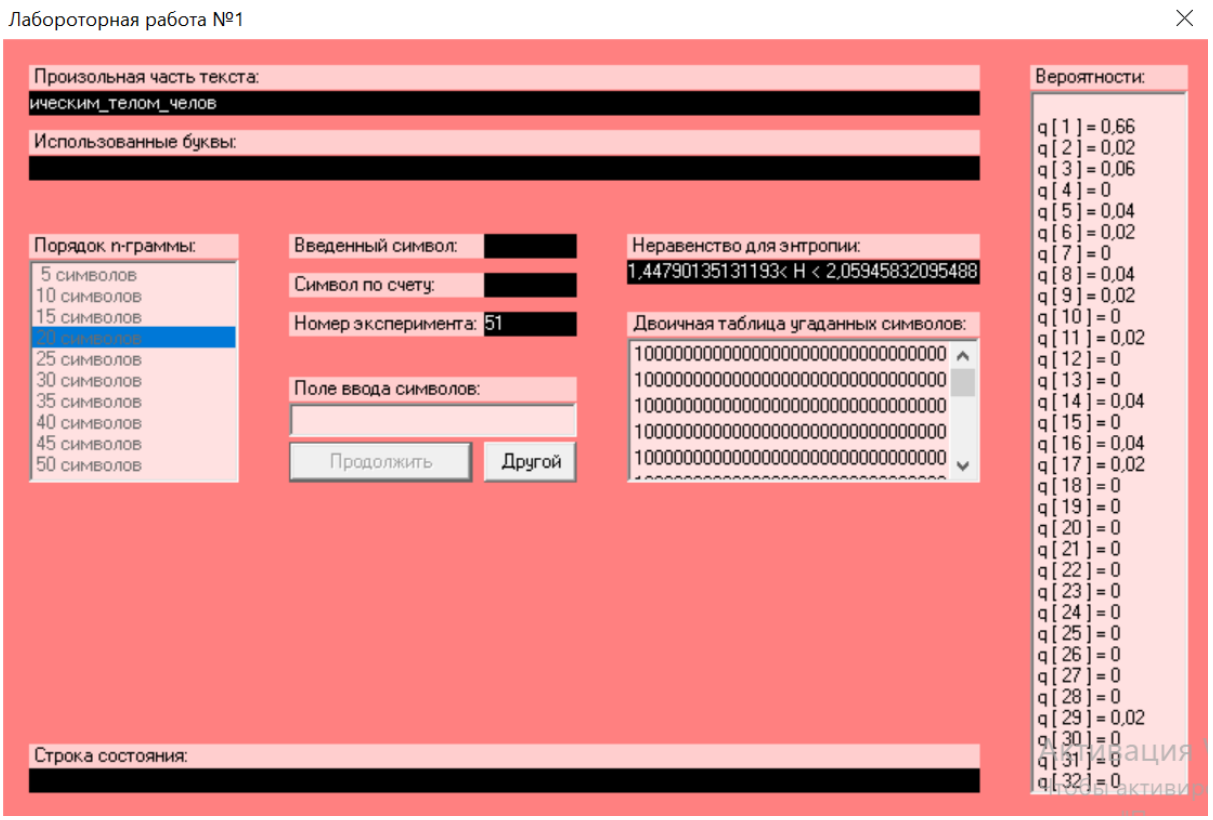
Другою частиною практикуму було дослідження значень H(10), H(20) та H(30)

H(10):



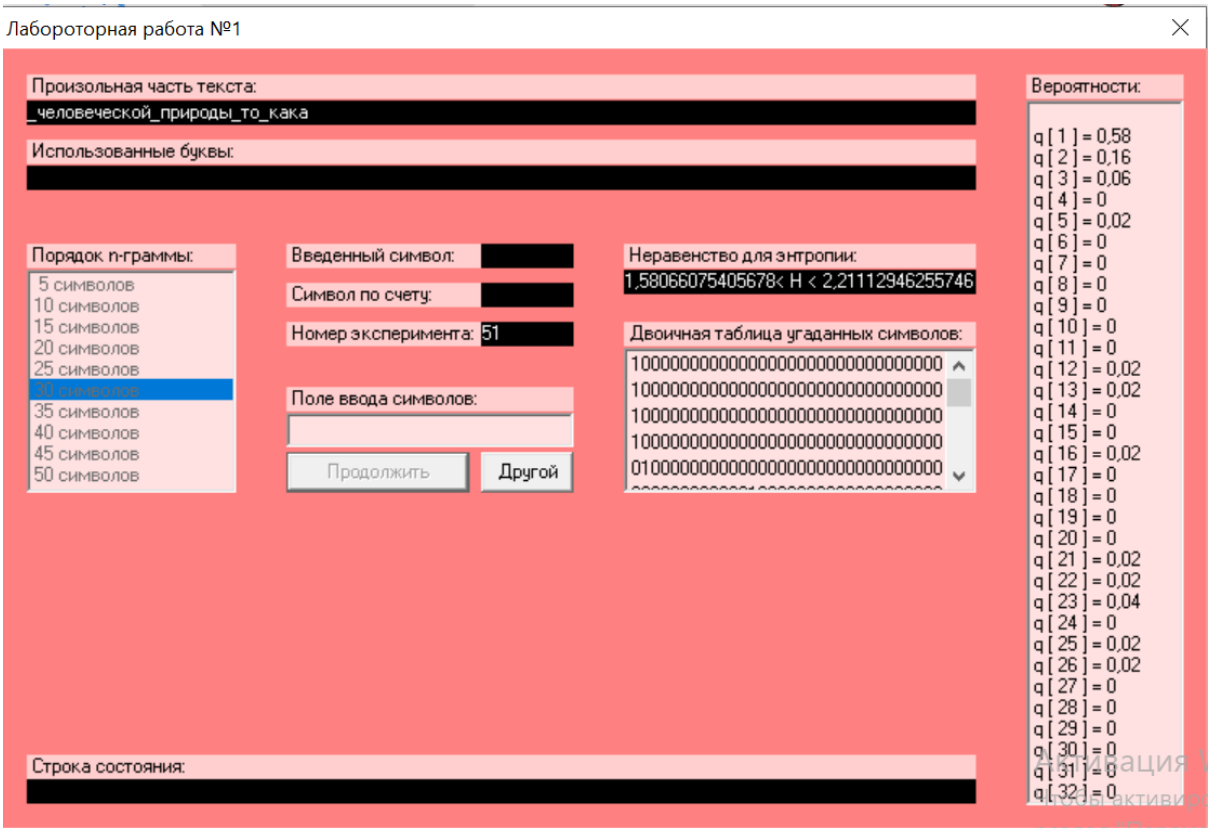
Надлишковість: 55.8% > R1 > 45%

H(20):



Надлишковість: 71% > R2 > 41%

H(30):



Надлишковість: 68% > R > 56%

Висновок:

Під час виконання даної лабораторної роботи мені довелося ознайомитися не лише з ентропією, дослідженнями надлишковості мови та створеннями масивів, а також майже з самого початку вивчити нову мову програмування. Також корисним було закріплення роботи з файлами та з гітхабом.