Міністерство освіти і науки України

Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

Криптографія

Лабораторна робота №1

Виконав студент групи ФБ-13

Лагно Костянтин

Київ 2023

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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.

Опис труднощів: серед труднощів, які виникли під час виконання даного практикуму, можу виділити такі:

- Проблеми з форматуванням тексту – прийняв рішення написати окрему програму для підготовки тексту для подальшої роботи, оскільки початкова версія пропускала певні спеціальні символи, які були в тексті, як-от спецсимвол перенесення рядка, тощо, і потм вони вспливали під час обрахунку біаграм.
- Недостатній рівень знання мови програмування – довелось переписувати програму декількома версіями.
- Криві ручки – невиправимо ;)

Хід роботи:

Відформатувавши текст так, як мені потрібно, приступив до самого підрахунку частот. Усі вихідні дані додатково додам окремою таблицею, тут наведу лише частину для загального розуміння.

1. Власна програма

1) Текст з пробілами, частота букв, їх ентропія H1 та надлишковість R1:

```
Ентропія тексту : 4.370842906710109
Надлишковість тексту : 0.1258314186579781
```

Літера ' '	0.16393981782151837
Літера 'о'	0.09023857498903709
Літера 'а'	0.07184587545031168
Літера 'е'	0.07078427780830739
Літера 'т'	0.053582232879829925
Літера 'и'	0.053054903332167655
Літера 'н'	0.051722702369652464
Літера 'с'	0.045758327643725054
Літера 'л'	0.04145504099293378
Літера 'в'	0.03591669303314405
Літера 'р'	0.0357515556221656
Літера 'к'	0.028173274938524475
Літера 'м'	0.027852714081919257
Літера 'д'	0.027064495179097765
Літера 'у'	0.024957952407120614

2) Текст без пробілів, частота букв, ентропія H1 та надлишковість R1:

```
Ентропія тексту : 4.458041299231585
Надлишковість тексту : 0.1001484358048268
```

Літера 'о'	0.10793310925764554
Літера 'а'	0.08593385617660484
Літера 'е'	0.08466409394580687
Літера 'т'	0.06408896634715133
Літера 'и'	0.06345823478152621
Літера 'н'	0.06186480766836798
Літера 'с'	0.05473090169716586
Літера 'л'	0.04958380015768289
Літера 'в'	0.04295945889870949
Літера 'р'	0.04276194032947425
Літера 'к'	0.03369766380347732
Літера 'м'	0.033314245404373624
Літера 'д'	0.03237146769575501
Літера 'у'	0.029851861073073572
Літера 'п'	0.029207850948172125

3) Текст з пробілами, біграми з перетином, частота, ентропія H2 та надлишковість R2:

```
Ентропія тексту : 3.956720339709514 біт
Надлишковість тексту : 0.20865593205809718
```

Біграма 'ау'	0.00010407834462699016
Біграма 'мь'	0.00010407834462699016
Біграма 'ху'	0.00010407834462699016
Біграма 'юш'	0.00010407834462699016
Біграма 'дю'	0.00010685376715037655
Біграма 'мс'	0.00010685376715037655
Біграма 'рх'	0.00010685376715037655
Біграма 'тд'	0.00010685376715037655
Біграма 'фи'	0.00010685376715037655
Біграма 'ищ'	0.00011101690093545616
Біграма 'ыг'	0.00011379232345884257
Біграма 'яю'	0.00011379232345884257
Біграма 'бс'	0.00011518003472053576
Біграма 'ге'	0.00011518003472053576
Біграма 'нр'	0.00011795545724392217

4) Текст без пробілів, біграми з перетином, частота, ентропія H2 та надлишковість R2:

```
Ентропія тексту : 4.13476908959826 біт
Надлишковість тексту : 0.16540063603669064
```

Біграма 'мх'	0.00010124918253733771
Біграма 'пн'	0.00010124918253733771
Біграма 'юж'	0.00010124918253733771
Біграма 'жк'	0.00010290900520188423
Біграма 'рз'	0.00010290900520188423
Біграма 'зь'	0.00010456882786643076
Біграма 'жч'	0.00010622865053097727
Біграма 'ыа'	0.00010622865053097727
Біграма 'юр'	0.00010622865053097727
Біграма 'хз'	0.00010788847319552379
Біграма 'йе'	0.00011120811852461683
Біграма 'рч'	0.00011120811852461683
Біграма 'нг'	0.00011286794118916335
Біграма 'цв'	0.00011286794118916335
Біграма 'дд'	0.00011618758651825638

5) Текст з пробілами, біграми без перетину, частота, ентропія H3 та надлишковість R3:

```
Ентропія тексту : 3.9563032855994447 біт
Надлишковість тексту : 0.2087393428801111
```

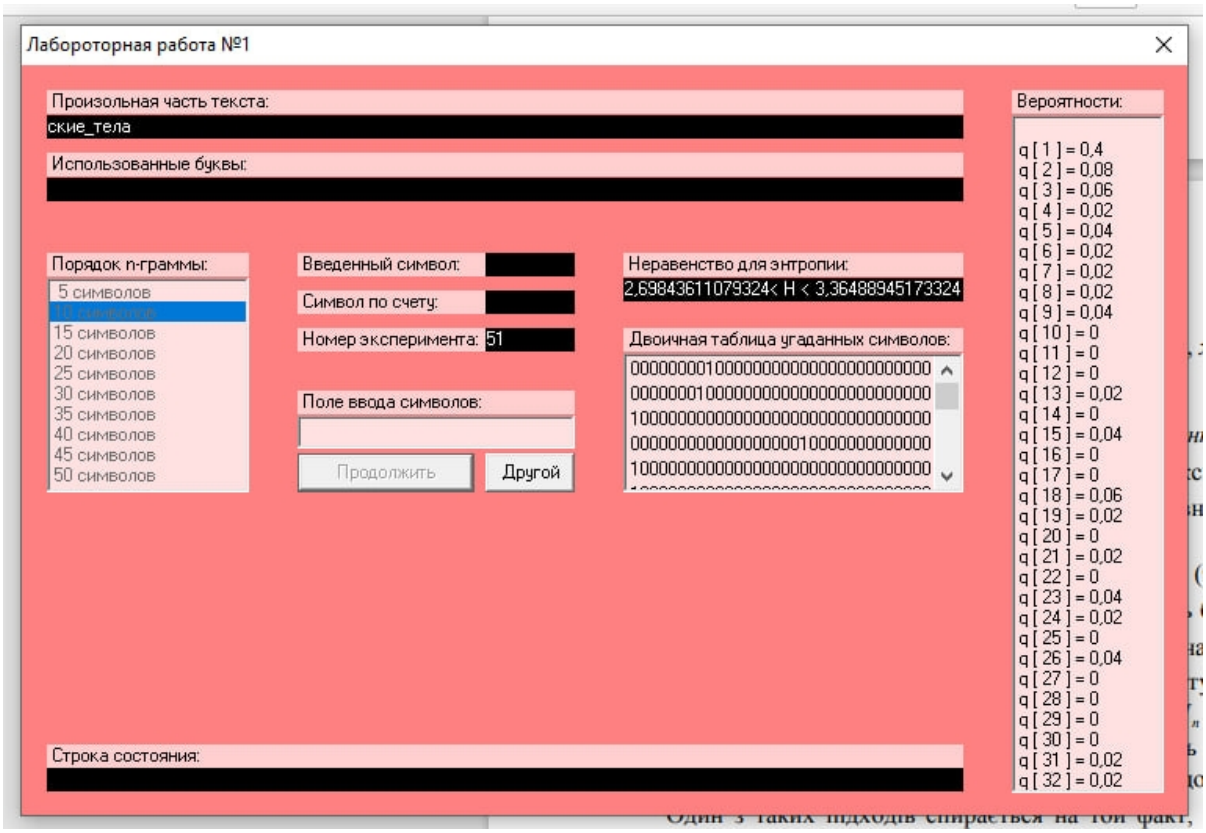
Біграма 2 'ях'	0.00010269049086054631
Біграма 2 'юш'	0.00010269049086054631
Біграма 2 'фи'	0.00010269049086054631
Біграма 2 'лч'	0.00010546590953245297
Біграма 2 'мс'	0.00010546590953245297
Біграма 2 'ац'	0.00010546590953245297
Біграма 2 'чо'	0.00010824132820435963
Біграма 2 'ищ'	0.00010824132820435963
Біграма 2 'ею'	0.00010824132820435963
Біграма 2 'нр'	0.00010824132820435963
Біграма 2 'дю'	0.00011101674687626629
Біграма 2 'цы'	0.00011101674687626629
Біграма 2 'ыг'	0.00011101674687626629
Біграма 2 'зб'	0.00011379216554817295
Біграма 2 'рх'	0.00011379216554817295

6) Текст без пробілів, біграми без перетину, частота, ентропія H3 та надлишковість R3:

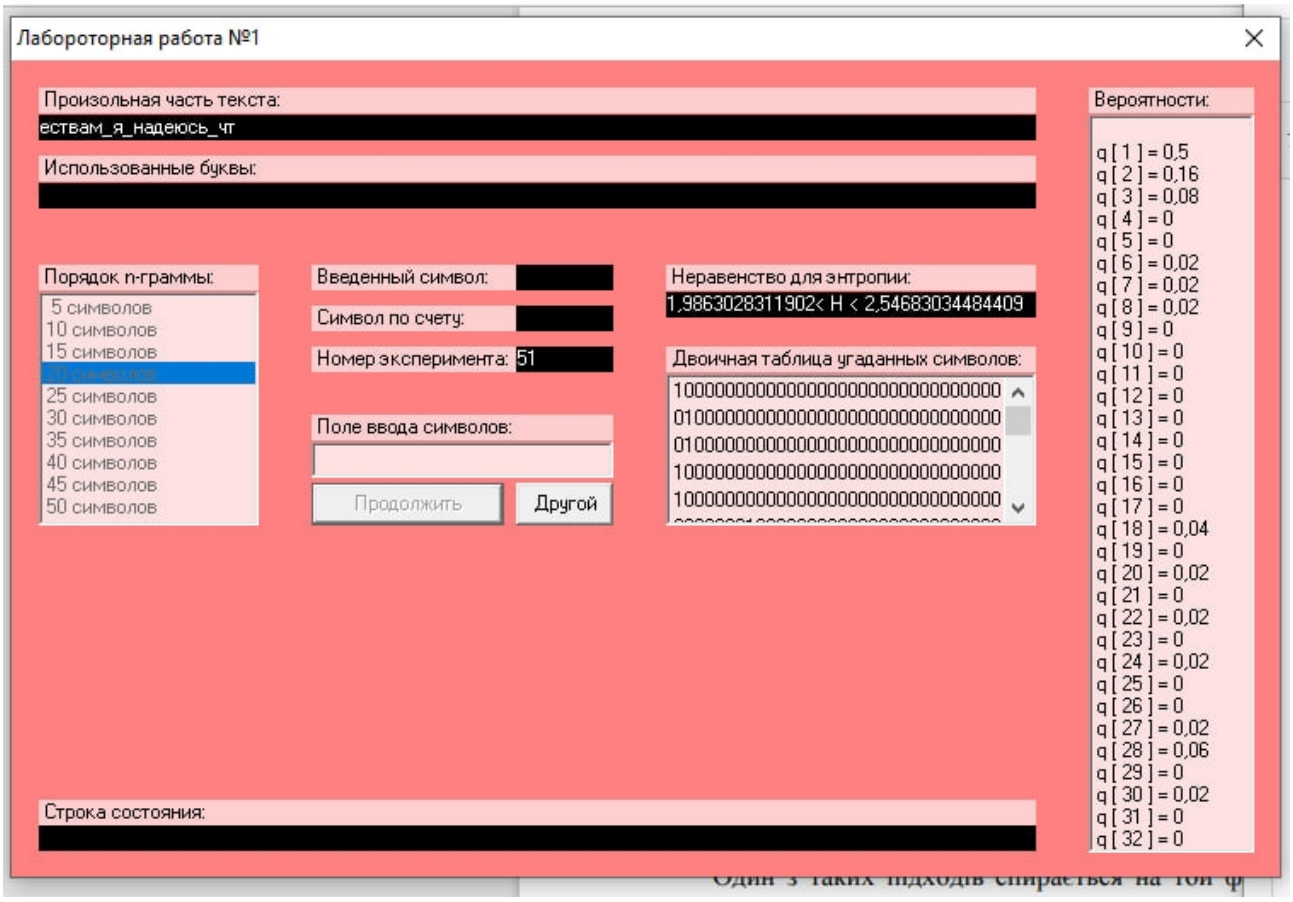
```
Ентропія тексту : 4.134948485088677 біт
Надлишковість тексту : 0.16536442522081296
```

Біграма 2 'гв'	0.00010290900520188423
Біграма 2 'пт'	0.00010290900520188423
Біграма 2 'фи'	0.00010290900520188423
Біграма 2 'зз'	0.00010622865053097727
Біграма 2 'сз'	0.00010622865053097727
Біграма 2 'сэ'	0.00010622865053097727
Біграма 2 'нг'	0.00010622865053097727
Біграма 2 'жч'	0.00010622865053097727
Біграма 2 'хт'	0.00010622865053097727
Біграма 2 'кц'	0.00010622865053097727
Біграма 2 'зь'	0.0001095482958600703
Біграма 2 'хз'	0.0001095482958600703
Біграма 2 'пк'	0.00011286794118916335
Біграма 2 'йе'	0.00011286794118916335
Біграма 2 'нм'	0.00011286794118916335

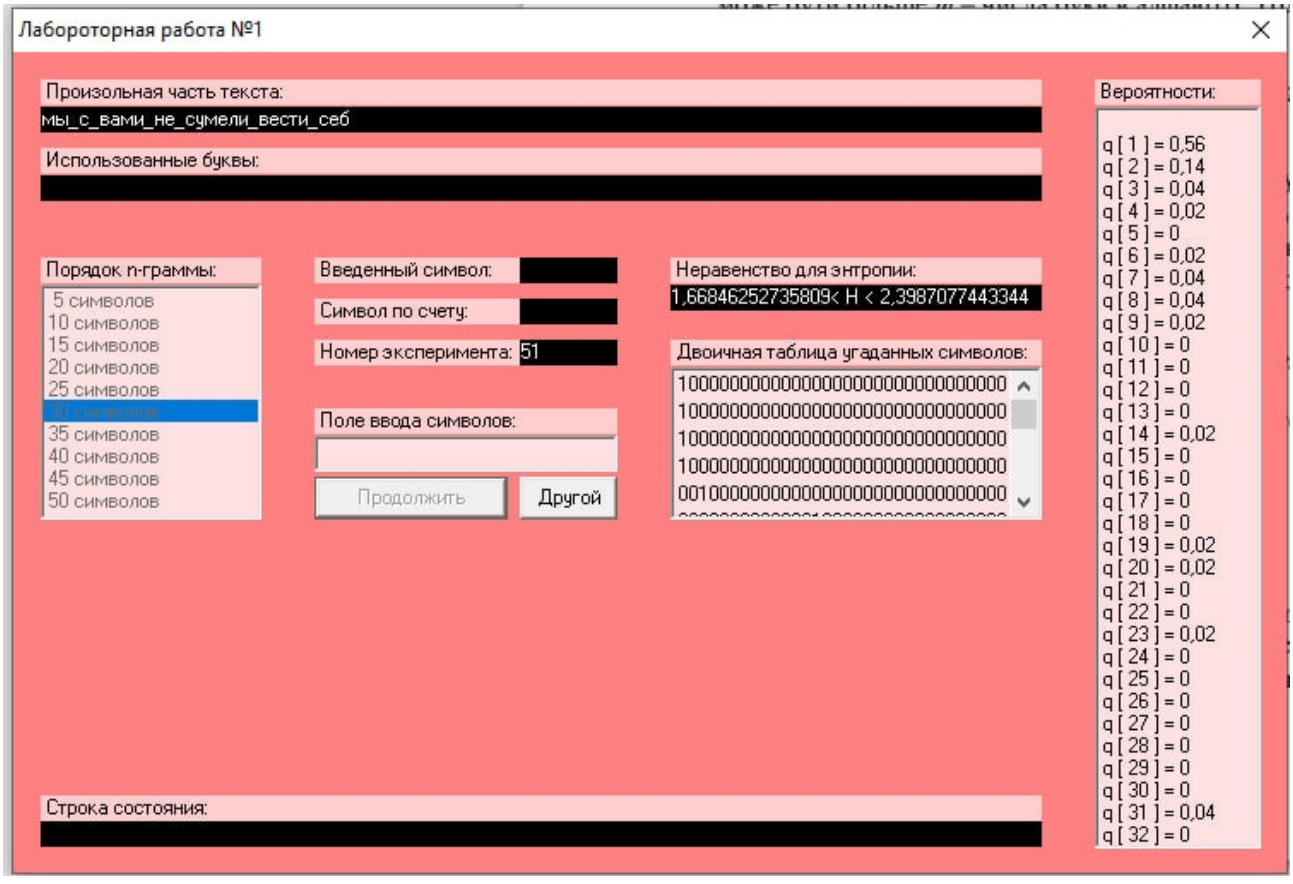
2. Оцінки H10, H20, H30 програмою CoolPinkProgram



H10


Лабороторная работа №1
Произольная часть текста:
еством_я_надеюсь_чт
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,9863028311902< H < 2,54683034484409
Двоичная таблица угаданных символов:
Вероятности:
q[1] = 0,5
q[2] = 0,16
q[3] = 0,08
q[4] = 0
q[5] = 0
q[6] = 0,02
q[7] = 0,02
q[8] = 0,02
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0,04
q[19] = 0
q[20] = 0,02
q[21] = 0
q[22] = 0,02
q[23] = 0
q[24] = 0,02
q[25] = 0
q[26] = 0
q[27] = 0,02
q[28] = 0,06
q[29] = 0
q[30] = 0,02
q[31] = 0
q[32] = 0
Строка состояния:

H20


Лабороторная работа №1
Произольная часть текста:
мы_с_вами_не_сумели_вести_себ
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,66846252735809< H < 2,3987077443344
Двоичная таблица угаданных символов:
Вероятности:
q[1] = 0,56
q[2] = 0,14
q[3] = 0,04
q[4] = 0,02
q[5] = 0
q[6] = 0,02
q[7] = 0,04
q[8] = 0,04
q[9] = 0,02
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0,02
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0
q[19] = 0,02
q[20] = 0,02
q[21] = 0
q[22] = 0
q[23] = 0,02
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0,04
q[32] = 0
Строка состояния:

H30

Таблиця ентропій для СРР:

	Найвище значення	Найнижче значення
H10	3.36488	2.69843
H20	2.54683	1.98630
H30	2.39870	1.66846

3. Оцінка надлишковості російської мови.

Формула: $R_n = 1 - \frac{H_n}{H0} = 1 - \frac{H_n}{\log_2 m}$

	Найнижче значення	Найвище значення
R10	0,327024	0,460314
R20	0,490634	0,60274
R30	0,52026	0,666308

Порівнюючи ці дані з даними, отриманими моєю програмою, можна зробити висновок, що при меншому значення n надлишковість зростає швидше, ніж при більших значеннях n. Це пов’язано з зниженням ентропії, оскільки в моїй програмі ентропія в середньому була в районі 4, при n = 2, а при використанні СРР і n > 10, ентропія зменшувалась.

Також можна сказати, що надлишковість російського тексту прямує до значення приблизно 0.74.

Висновки: у ході виконання лабораторної роботи я дізнався детальніше про ентропію та надлишіковість в теорії інформації, використав ці знання під час практичних дослідів.