

Міністерство освіти і науки України

Національний технічний університет України

"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

Комп’ютерний практикум

Робота № 1

Виконали

студенти гр. ФБ-06,

Зінов'єв Андрій, Даценко Валерія

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 комп'ютерного практикуму

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 H(10), H(20), H(30).

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Під час виконання даної лабораторної роботи використовувався написаний нами код на Python 3, що збережений у файлі *1lab.py* та текст у файлі *thetext.txt*.

Код складається з декількох функцій:

1. Функція *adding()* - рахує кількість повторів кожного символу;
2. Функція *adding\_bigram\_list()* - аналогічно до 1, але для біграм;
3. Функція *replace()* - конвертує масив, створений функцією *adding()* у новий масив зі значеннями частот. Іншими словами, створює наші пари - ансамблі;
4. Функція *replace\_bigrams()* - конвертує вихідний масив функції *adding\_bigram\_list()* у новий, за принципом функції *replace();*
5. Функція *count\_entropy\_value()* - допомогає обчислити значення ентропії імовірнісного ансамблю *H(Z)*;
6. Функція *count\_redundancy() -* повертає значення надлишковості джерела відкритого тексту без втрачання його змісту, величину *R*.

При розробці цього рішення ми зіткнулись з певними проблемами у написанні та використанні деяких алгоритмів та функцій. Такими як, читання тексту з текстового файлу: проблема з кодуванням та труднощі з видалянням повторюваних пробілів у тексті. Була проблема знайти імовірності для ансамблів, а саме взаємодія з типом даних *dict -* його особливостями перезапису значень. Завдяки бібліотеки *pandas* не обійшлось без перешкод, спочатку записували дані у різні файли, але тільки потім з декількох зробили один. Чимало часу пішло на те, щоб зрозуміти принцип роботи програми *CoolPinkProgram*, бувало не раз, що експеримент завершався чарівною кнопкою *"Другой".*

Нижче приведені таблиці з результатами виконання програми

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Symbol | Chance |  | Symbol | Chance |
|  | 0.16336 |  | г | 0.01687 |
| о | 0.09484 |  | ь | 0.01677 |
| а | 0.07073 |  | ы | 0.01592 |
| е | 0.06869 |  | б | 0.01484 |
| и | 0.05476 |  | з | 0.01454 |
| н | 0.05463 |  | ч | 0.01190 |
| т | 0.04836 |  | й | 0.00969 |
| с | 0.04414 |  | ж | 0.00871 |
| л | 0.04204 |  | ш | 0.00815 |
| в | 0.03816 |  | х | 0.00722 |
| р | 0.03703 |  | ю | 0.00548 |
| к | 0.02886 |  | ц | 0.00296 |
| д | 0.02558 |  | э | 0.00260 |
| м | 0.02526 |  | щ | 0.00241 |
| у | 0.02327 |  | ф | 0.00161 |
| п | 0.02117 |  | ъ | 0.00038 |
| я | 0.01902 |  | ё | 0.00004 |

*Таблиця ансамблів літер і ймовірностей у тексті з пробілами.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Symbol | Chance |  | Symbol | Chance |
| о | 0.11336 |  | ь | 0.02004 |
| а | 0.08454 |  | ы | 0.01903 |
| е | 0.08211 |  | б | 0.01774 |
| и | 0.06545 |  | з | 0.01738 |
| н | 0.06530 |  | ч | 0.01423 |
| т | 0.05781 |  | й | 0.01158 |
| с | 0.05275 |  | ж | 0.01041 |
| л | 0.05025 |  | ш | 0.00974 |
| в | 0.04561 |  | х | 0.00863 |
| р | 0.04426 |  | ю | 0.00655 |
| к | 0.03449 |  | ц | 0.00353 |
| д | 0.03057 |  | э | 0.00311 |
| м | 0.03019 |  | щ | 0.00288 |
| у | 0.02781 |  | ф | 0.00192 |
| п | 0.02531 |  | ъ | 0.00045 |
| я | 0.02273 |  | ё | 0.00005 |
| г | 0.02017 |  |  |  |

*Таблиця ансамблів літер і ймовірностей у тексті без пробілів.*

|  |  |
| --- | --- |
| Bigram | Chance |
| о | 0.02114 |
| и | 0.01809 |
| а | 0.01747 |
| е | 0.01682 |
| с | 0.01605 |
| н | 0.01549 |
| п | 0.01512 |
| в | 0.014887 |
| то | 0.013673 |
| о | 0.012 |

*Таблиця ансамблів біграм і ймовірностей з пробілами.*

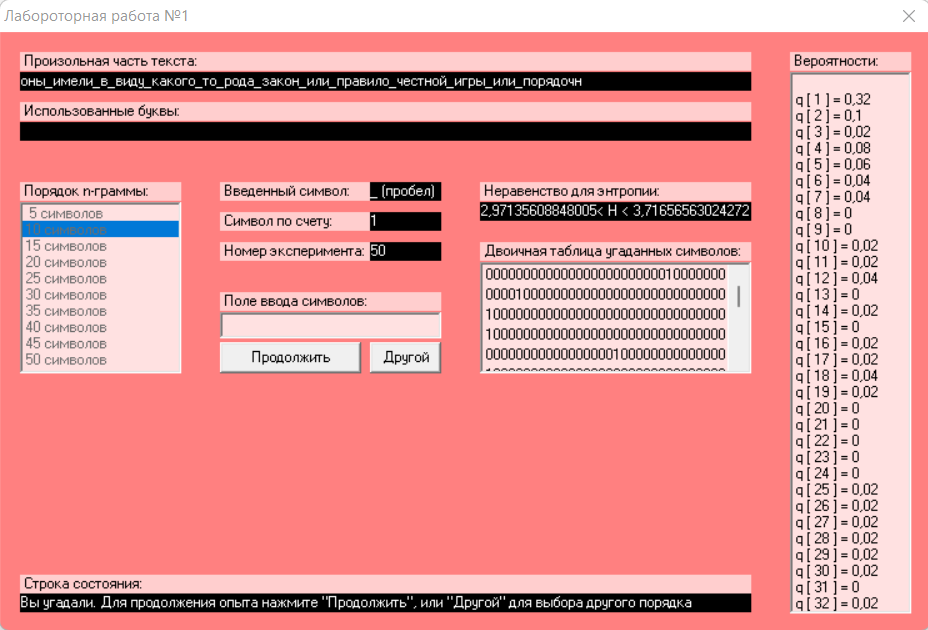
|  |  |
| --- | --- |
| Bigram | Chance |
| то | 0.01665 |
| ст | 0.01305 |
| на | 0.01282 |
| ов | 0.01161 |
| ал | 0.01085 |
| не | 0.01085 |
| го | 0.01073 |
| он | 0.01041 |
| ос | 0.01018 |
| ко | 0.01016 |

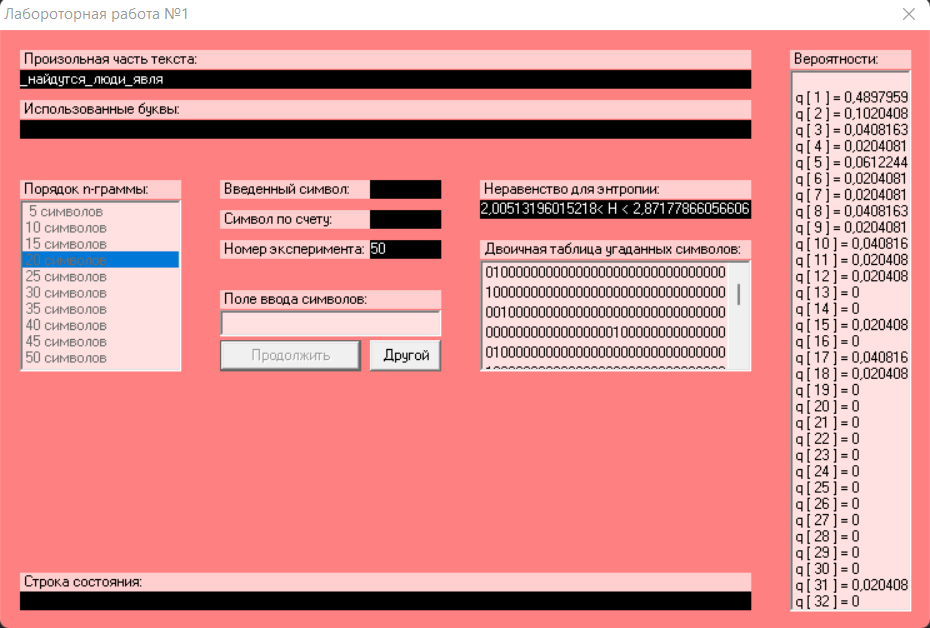
*Таблиця ансамблів біграм і ймовірностей без пробілів.*

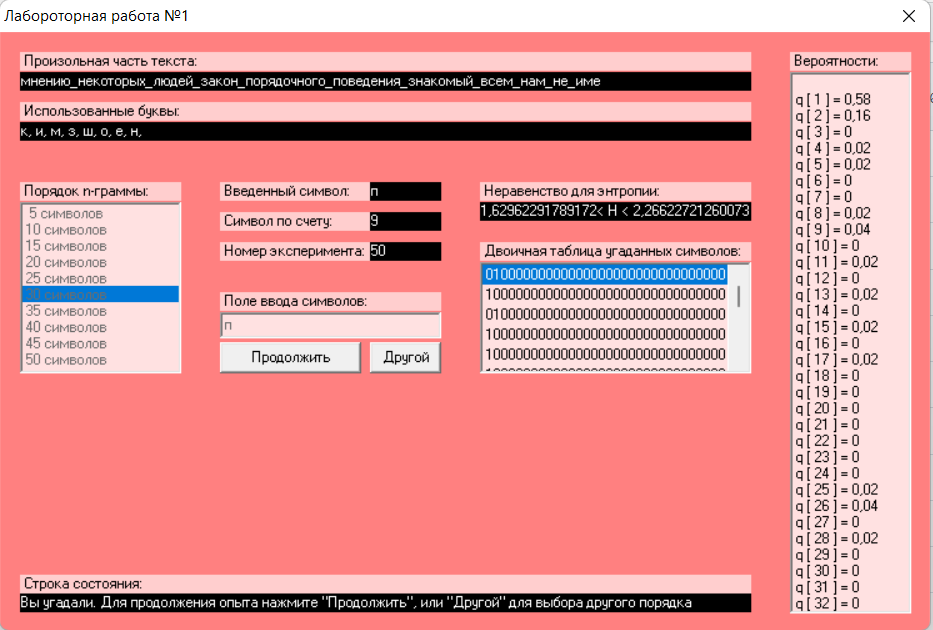
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Величина | Символи | | Біграми | |
| З пробілами | Без пробілів | З пробілами | Без пробілів |
| H | 4.3812405 | 4.4689349 | 3.9696057 | 4.1481197 |
|  |
|  |
| R | 0.13146347 | 0.11407896 | 0.21306591 | 0.17767731 |  |
|  |
|  |

*Таблиця значень ентропій і надлишковостей для усіх випадків H1, H2.*

CoolPinkProgram







Отримали результати для H(10) H(20) H(30)

2.97135< H(10) < 3.71656

2.00513 < H(20) < 2.87177

1.62962 < H(30) < 2.26622

|  |  |  |  |
| --- | --- | --- | --- |
| CoolPinkProgram | | | |
| Величина | H(10) | H(20) | H(30) |
|  |
| H | 3.71656 | 2.87177 | 2.26622 |  |
|  |
| R | 0.26322 | 0.4307 | 0.55074 |  |
|  |

*Значення ентропії та надлишковості для випадків H(10), H(20), H(30).*

Висновки:

У ході даної лабораторної роботи ми ознайомились та засвоїли поняття ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набули практичних навичок щодо оцінки ентропії на символ джерела. При виконанні даного практикуму ми вдосконалили навички програмування та узагальнили розуміння криптографічних визначень. Визначили, що у російській мові (так само як і в українській) найпоширенішими літерами є: о, а, е, и, н, т, с, л.