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 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

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 видаляннням 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
	0.16336
о	0.09484
а	0.07073
е	0.06869
и	0.05476
н	0.05463
т	0.04836
с	0.04414
л	0.04204
в	0.03816
р	0.03703
к	0.02886
д	0.02558
м	0.02526
у	0.02327
п	0.02117
я	0.01902

Symbol	Chance
г	0.01687
ь	0.01677
ы	0.01592
б	0.01484
з	0.01454
ч	0.01190
й	0.00969
ж	0.00871
ш	0.00815
х	0.00722
ю	0.00548
ц	0.00296
э	0.00260
щ	0.00241
ф	0.00161
ъ	0.00038
ё	0.00004

Таблиця ансамблів літер і ймовірностей у тексті з пробілами.

Symbol	Chance
о	0.11336
а	0.08454
е	0.08211
и	0.06545
н	0.06530
т	0.05781
с	0.05275
л	0.05025
в	0.04561
р	0.04426
к	0.03449
д	0.03057
м	0.03019
у	0.02781
п	0.02531
я	0.02273
г	0.02017

Symbol	Chance
ь	0.02004
ы	0.01903
б	0.01774
з	0.01738
ч	0.01423
й	0.01158
ж	0.01041
ш	0.00974
х	0.00863
ю	0.00655
ц	0.00353
э	0.00311
щ	0.00288
ф	0.00192
ъ	0.00045
ё	0.00005

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

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 з пробілами та без (перехресно).

Bigram	Chance
о	0.02121
и	0.01820
а	0.01751
е	0.01671
с	0.01586
н	0.01558
п	0.01511
в	0.01485
то	0.01365
о	0.01192

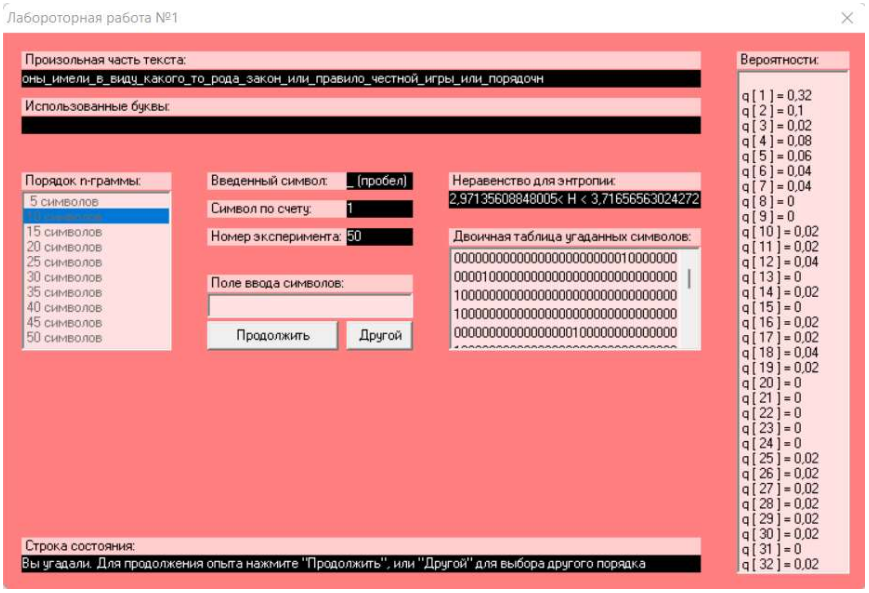
Bigram	Chance
то	0.01653
ст	0.01311
на	0.01284
ов	0.01149
ал	0.01100
не	0.01074
го	0.01069
он	0.01049
но	0.01019
ос	0.01018

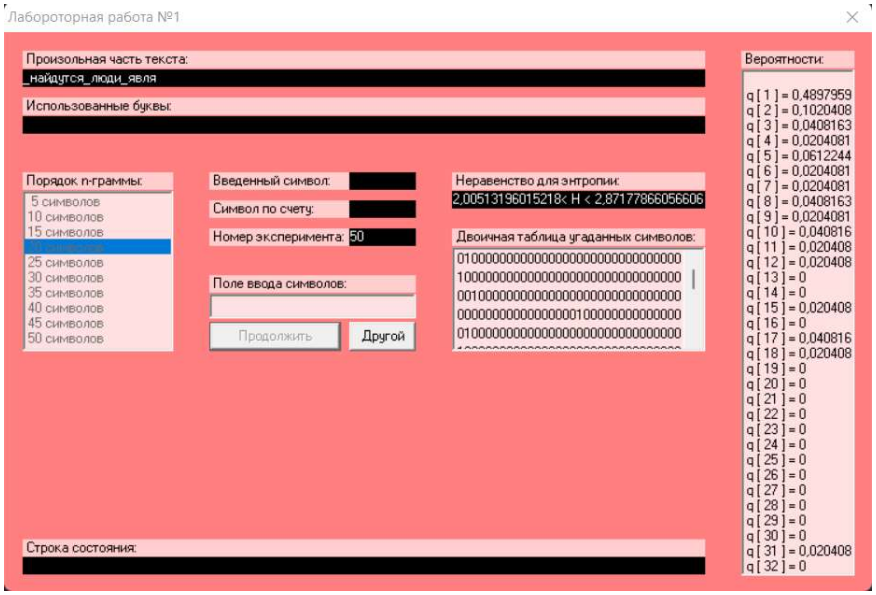
Таблиця ансамблів біграм і ймовірностей з пробілами та без (не перехресно).

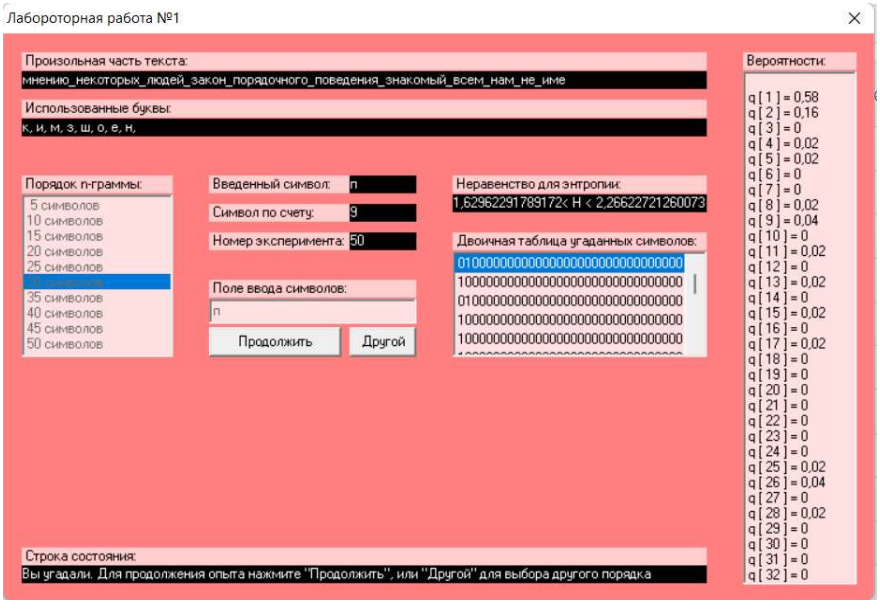
Величина	Символи		Біграми	
	З пробілами	Без пробілів	З пробілами	Без пробілів
H	4.3812405	4.4689349	3.9696057	4.1481197
R	0.13146347	0.11407896	0.21306591	0.17767731

Таблиця значень ентропій і надлишковостей для усіх випадків H_1, H_2.

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.