Міністерство освіти і науки України Національний технічний університет України “Київський політехнічний інститут ім. Ігоря Сікорського” Фізико-технічний інститут

КРИПТОГРАФІЯ КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: Студент групи ФБ-05 Даниленко Данило,
Студентка ФБ-05 Мірошніченко Ілона

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 1 H та 2 H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 H та 2 H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1 H та 2 H 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

Перед виконанням роботи були розглянуті теоретичні відомості в методичних вказівках. В якості експериментального тексту була взята книга «Good Omens» в перекладі на російську мову. Оригінал тексту можна знайти у файлі badtxt.txt. Відредагований текст з пробілами міститься у файлі spaces.txt, а без пробілів у nospaces.txt. В ході виконання роботи було прийнято рішення пробіли замінити на «_» для кращого сприйняття. Усі таблиці також наведені у файлах з відповідними назвами. Єдина відмінність - таблиці біграм були виконанні у двох варіантах(таблиці у вигляді матриці і звичайні)

Monograms with space

h1 with space - 4.4769572317595205

r1 with space - 0.1200019790888095

о - 0.102851	у - 0.035314	х - 0.011443
а - 0.088136	м - 0.029008	й - 0.010844
е - 0.078324	д - 0.028709	ж - 0.010705
и - 0.075033	п - 0.026668	ш - 0.00962
н - 0.065523	я - 0.021505	ю - 0.005917
т - 0.062623	ы - 0.018928	щ - 0.004627
с - 0.050766	ь - 0.018739	ц - 0.004129
р - 0.048078	б - 0.018396	э - 0.003246
л - 0.042801	з - 0.017728	ф - 0.00091
в - 0.042185	г - 0.015919	ъ - 0.00018
к - 0.036785	ч - 0.014189	ё - 0.00017

Monograms without space

h1 with space - 4.4769572317595205

r1 without space - 0.11248977413377115

о - 0.102851	у - 0.035314	х - 0.011443
а - 0.088136	м - 0.029008	й - 0.010844
е - 0.078324	д - 0.028709	ж - 0.010705
и - 0.075033	п - 0.026668	ш - 0.00962
н - 0.065523	я - 0.021505	ю - 0.005917
т - 0.062623	ы - 0.018928	щ - 0.004627
с - 0.050766	ь - 0.018739	ц - 0.004129
р - 0.048078	б - 0.018396	э - 0.003246
л - 0.042801	з - 0.017728	ф - 0.00091
в - 0.042185	г - 0.015919	ъ - 0.00018
к - 0.036785	ч - 0.014189	ё - 0.00017

Bigrams with space

h2 with space - 4.112855148278171

r2 with space - 0.19157047891727508

то - 0.016059	по - 0.011972	ко - 0.009658
но - 0.015962	не - 0.011686	ал - 0.009501
на - 0.013675	ка - 0.010931	ть - 0.009323
ст - 0.013129	ни - 0.010508	от - 0.009215
ра - 0.012474	ро - 0.009712	та - 0.009139

Bigrams without space

h2 without space - 4.112855148278171

r2 without space - 0.16571603296123505

то - 0.013936	по - 0.010055	ос - 0.009124
но - 0.013602	не - 0.009863	от - 0.008944
на - 0.011513	ен - 0.009316	ов - 0.008794
ст - 0.011207	ка - 0.009222	он - 0.008491
ра - 0.010488	ни - 0.009173	ак - 0.008325

h2 without space without intersection - 4.162934395360695

r2 without space without intersection - 0.1747404550756756

h2 with space without intersection - 3.9970781337680905

r2 with space without intersection - 0.21432779786441958

Bigrams with spaces without intersection

то - 0.016965	по - 0.012647	ко - 0.010204
но - 0.016864	не - 0.012346	ал - 0.010037
на - 0.014447	ка - 0.011549	ть - 0.009849
ст - 0.01387	ни - 0.011102	от - 0.009735
ра - 0.013178	ро - 0.01026	та - 0.009655

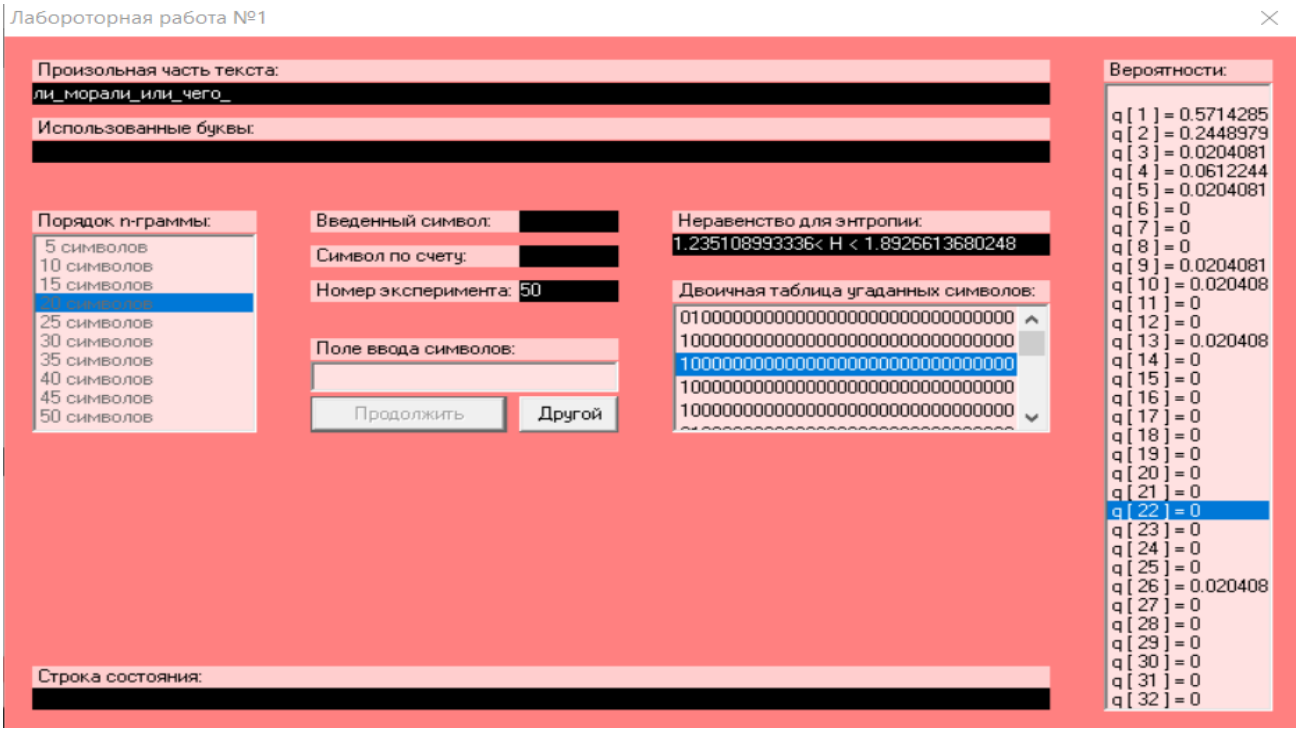
Bigrams without space without intersection

то - 0.014266	по - 0.010293	ос - 0.00934
но - 0.013924	не - 0.010097	от - 0.009156
на - 0.011785	ен - 0.009536	ов - 0.009003
ст - 0.011472	ка - 0.00944	он - 0.008692
ра - 0.010736	ни - 0.00939	ак - 0.008522

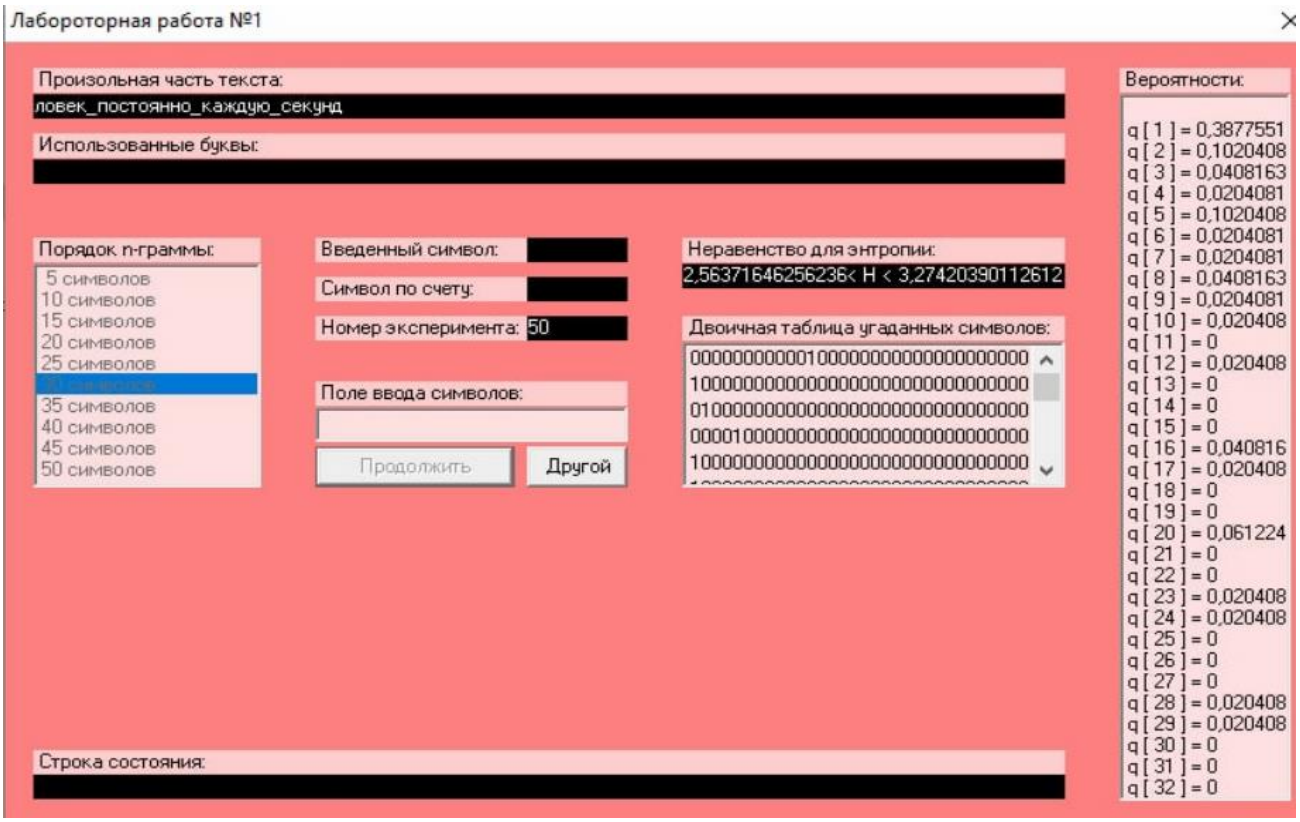
Результати експериментів у CoolPinkProgram

n=10:

n=20:


Лабороторная работа №1
Произольная часть текста:
ли_морали_или_чего_
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: 50
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1.235108993336< H < 1.8926613680248
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0.5714285
q [2] = 0.2448979
q [3] = 0.0204081
q [4] = 0.0612244
q [5] = 0.0204081
q [6] = 0
q [7] = 0
q [8] = 0
q [9] = 0.0204081
q [10] = 0.020408
q [11] = 0
q [12] = 0
q [13] = 0.020408
q [14] = 0
q [15] = 0
q [16] = 0
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0.020408
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0

n=30:


Лабороторная работа №1
Произольная часть текста:
ловек_постоянно_каждую_секунд
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: 50
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2,56371646256236< H < 3,27420390112612
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,3877551
q [2] = 0,1020408
q [3] = 0,0408163
q [4] = 0,0204081
q [5] = 0,1020408
q [6] = 0,0204081
q [7] = 0,0204081
q [8] = 0,0408163
q [9] = 0,0204081
q [10] = 0,020408
q [11] = 0
q [12] = 0,020408
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0,040816
q [17] = 0,020408
q [18] = 0
q [19] = 0
q [20] = 0,061224
q [21] = 0
q [22] = 0
q [23] = 0,020408
q [24] = 0,020408
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0,020408
q [29] = 0,020408
q [30] = 0
q [31] = 0
q [32] = 0

Висновки

Під час виконання даної лабораторної роботи, ми здобули навички з аналізу тексту. Використовуючи regex нам вдалося в досить простій формі отримати працююче рішення. Крім того, нами було засвоєно визначення ентропії та надлишковості, які ми інтегрували в наш розв’язок. В результаті ми перевірили статистику тексту за допомогою CoolPinkProgram. На нашу думку, дані навички стануть у нагоді у подальшому розвитку.