МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Компьютерний практикум №1

З дисципліни: «Криптографія»

Виконав:
Студент гр. ФБ-03
Гузенков А.М.
Перевірив:
Чорний О.М.

Київ – 2022

Тема

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.

Постановка задачі

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

Завдання 1

Для виконання першого завдання був написаний код (див. файл main.cpp), який підраховує частоту літер та біграм у тексті та рахує ентропію.

Таблиця з частотою літер:

Літера	Частота	Ймовірність
а	13 536	0,0782071
б	3 029	0,0175007
в	8 379	0,0484114
г	3 362	0,0194247
д	4 858	0,0280681
е	14 979	0,0865443
ж	1 841	0,0106368
з	3 015	0,0174198
и	12 162	0,0702685
й	2 419	0,0139763
к	5 533	0,0319681
л	8 060	0,0465683
м	5 229	0,0302116
н	12 496	0,0721982
о	19 720	0,113936
п	4 362	0,0252024
р	7 552	0,0436333
с	9 552	0,0551887
т	9 920	0,0573149
у	4 448	0,0256992
ф	284	0,00164087
х	1 644	0,00949855
ц	661	0,00381907
ч	2 349	0,0135718
ш	1 213	0,00700836
щ	576	0,00332796
ъ	46	0,000265775

ы	3 270	0,0188931
ь	3 109	0,0179629
э	522	0,00301596
ю	1 190	0,00687547
я	3 760	0,0217242
ё	3	0,0000173331

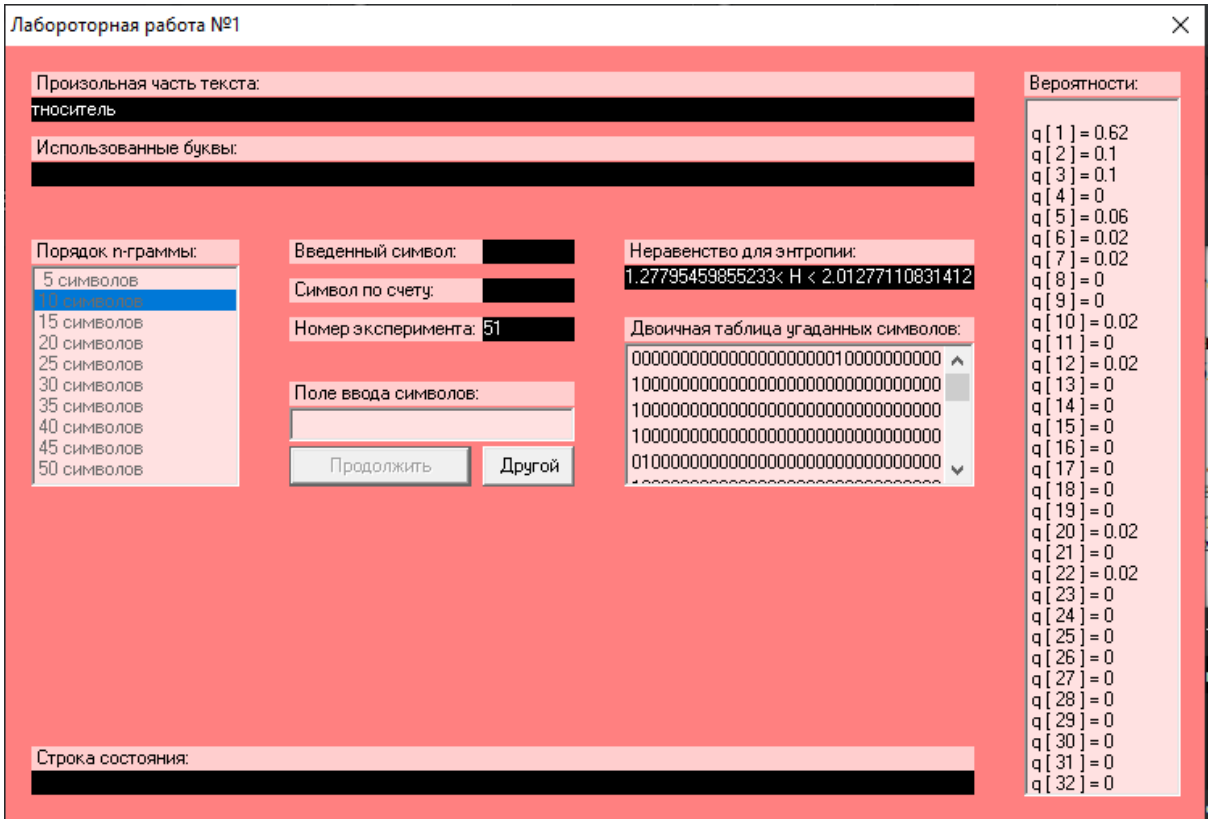
Таблиця з частотою біграм без пробілів та без перетинів (перші 10):

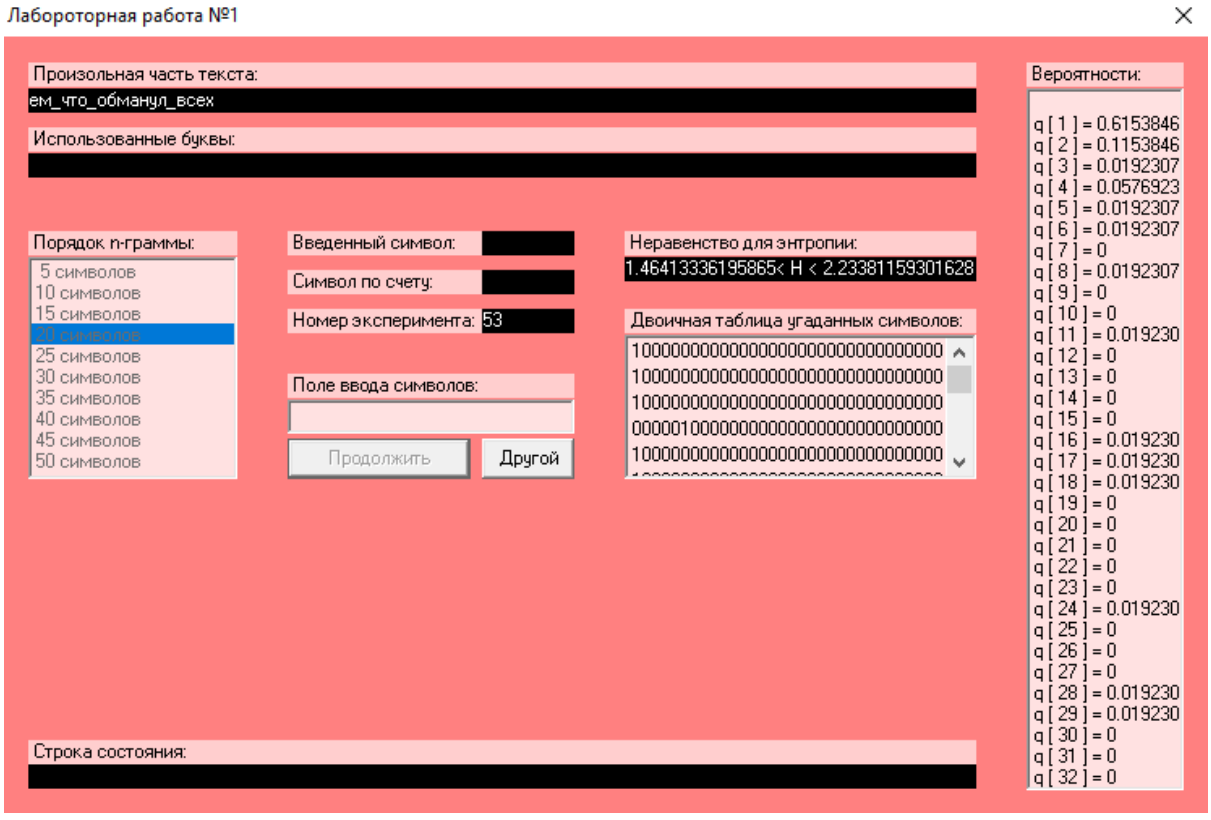
Біграма	Частота	Ймовірність
аа	38	0,000439103
аб	132	0,00152531
ав	511	0,00590478
аг	102	0,00117865
ад	165	0,00190663
ае	141	0,0016293
аж	185	0,00213774
аз	430	0,0049688
аи	144	0,00166397
ай	46	0,000531546

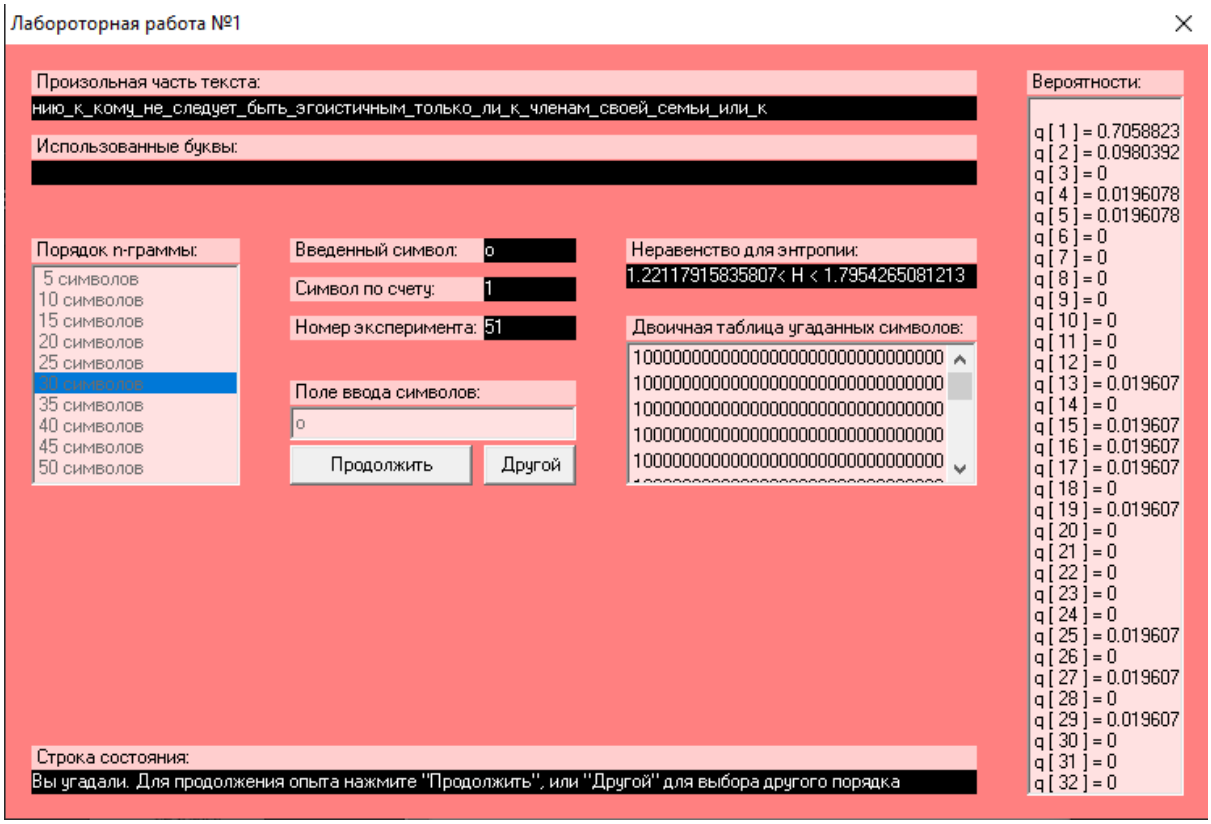
Переглянути повні таблиці частот біграм та монограм можна у файлах
table_bigrams_nospaces_cross.csv — біграми без пробілів та з перетинами
table_bigrams_nospaces_nocross.csv — біграми без пробілів та без перетинів
table_bigrams_spaces_cross.csv — біграми з пробілами та з перетинами
table_bigrams_spaces_nocross.csv — біграми з пробілами та без перетинів
table_letters_nospaces.csv — монограми без пробілів
table_letters_spaces.csv — монограми без пробілів

Результат роботи програми можна побачити нижче
Ентропія монограм без пробілів: 4,45249
Надлишковість для тексту без пробілів: 0,109503
Ентропія монограм з пробілами: 4,38827
Надлишковість для тексту з пробілами: 0,122346
Ентропія біграм без пробілів та з перетинами: 4,12868
Надлишковість для джерела біграм без пробілів та з перетинами: 0,174265
Ентропія біграм без пробілів та без перетинів: 4,10195
Надлишковість для джерела біграм без пробілів та без перетинів: 0,179611
Ентропія біграм з пробілами та перетинами: 3,97272
Надлишковість для джерела біграм з пробілами та перетинами: 0,205457
Ентропія біграм з пробілами та без перетинів: 3,95689
Надлишковість для джерела біграм з пробілами та без перетинів: 0,208623

Завдання 2:


Лабороторная работа №1
Произольная часть текста:
тноситель
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:
1.27795459855233< H < 2.01277110831412
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q[1] = 0.62
q[2] = 0.1
q[3] = 0.1
q[4] = 0
q[5] = 0.06
q[6] = 0.02
q[7] = 0.02
q[8] = 0
q[9] = 0
q[10] = 0.02
q[11] = 0
q[12] = 0.02
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0
q[19] = 0
q[20] = 0.02
q[21] = 0
q[22] = 0.02
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0


Лабороторная работа №1
Произольная часть текста:
ем_что_обманул_всех
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: 53
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1.46413336195865< H < 2.23381159301628
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q[1] = 0.6153846
q[2] = 0.1153846
q[3] = 0.0192307
q[4] = 0.0576923
q[5] = 0.0192307
q[6] = 0.0192307
q[7] = 0
q[8] = 0.0192307
q[9] = 0
q[10] = 0
q[11] = 0.019230
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0.019230
q[17] = 0.019230
q[18] = 0.019230
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0.019230
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0.019230
q[29] = 0.019230
q[30] = 0
q[31] = 0
q[32] = 0



Обрахуємо надлишковість для $H^{(10)}$, $H^{(20)}$, $H^{(30)}$:

$R = 1 - \frac{H_\infty}{H_0}$

$H_0 = \log_2 m = \log_2 32 = 5$

$H^{(10)} \approx 1.64536$
$R^{(10)} \approx 0.670928$
$H^{(20)} \approx 1.89897$
$R^{(20)} \approx 0.620206$
$H^{(30)} \approx 1.508305$
$R^{(30)} \approx 0{,}698339$

Труднощі, що виникли під час виконання практикуму

В ході роботи зтикнувся з труднощами кодування вводу-виводу. Рішенням цієї проблеми стало використання строк розширених (sizeof(wchar_t) = 4 байта у компіляторі GNU) символів та встановлення локалі ru_RU.UTF8 для всіх потокових об’єктів.