Міністерство освіти і науки України

Національний технічний університет України

"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

студент гр.ФБ-01 Заріцький О.В.

студент гр.ФБ-01 Свірщук Я.Ю.

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.

Постановка задачі

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

Для написання та виконання коду на python3 ми використовували Spyder IDE з пакету Anaconda.

Код програми

Lab1.py

```
import re

import collections

import pandas as pd

import math

import numpy as np


#--- Очищення тексту від непотрібних символів ---


print("/// Тест з пробілами чи без? ///")

userInput = input("+ з пробілами, - без пробілів: ")


file = open("./lab1.TXT").read()

file = file.lower()

file = file.replace("\n"," ")
```

```
if userInput == "+":
    clearString = re.sub(r'[^\w\s]+|[\d]+|_+', '',file).strip()
else:
    clearString = re.sub(r'[\W\s]+|[\d]+|_+', '',file).strip()

letter = sorted(clearString)
print(clearString)

#--- Кількість та частота появи літер ---

alphabet = []
for i in range(0,len(letter)):
    alphabet.append(letter[i])

print("\n/// Словник, де ключ - це літера, а значення - кількість цієї літери в тексті ///")
alphabetDict = dict(collections.Counter(alphabet))
print(alphabetDict)
print("\n/// Частота появи кожної літери алфавіту в тексті ///")
frequency = {k: alphabetDict[k] / len(letter) for k in alphabetDict}
print(frequency)

#--- Вивід кількості літер у вигляді датафрейму ---

alphabetFiltred = sorted(alphabetDict, key=lambda x : alphabetDict[x], reverse=1)

letterAmount = []
for i in range(0,len(alphabetFiltred)):
    letterAmount.append(alphabetDict[alphabetFiltred[i]])

df = pd.DataFrame(index = alphabetFiltred)
df['Кількість у тексті'] = letterAmount
df = df.rename(index={" ": "пробіл"})
```

```
print("\n", df.head(10))


#--- Вивід частоти появи літер ---

frequencyFiltred = sorted(frequency, key=lambda x : alphabetDict[x], reverse=1)

letterFrequency = []
for i in range(0,len(alphabetFiltred)):
    letterFrequency.append(frequency[alphabetFiltred[i]])

df2 = pd.DataFrame(index = alphabetFiltred)
df2['Частота'] = letterFrequency
df2 = df2.rename(index={" ": "пробіл"})
print("\n", df2.head(10))


#--- Пошук біграм, підрахунок їх кількості та частоти появи ---

file = clearString
bigram = []
bigramUncrossed = []
for i in range(0, len(file)-1):
    bigram.append(file[i]+file[i+1])

for i in range(0, len(file)-2, 2):
    bigramUncrossed.append(file[i]+file[i+1])


bigramAmount = dict(collections.Counter(bigram))

bigramFrequency = {k: bigramAmount[k] / len(bigram) for k in bigramAmount}
```

```
bigramUncrossedAmount = dict(collections.Counter(bigramUncrossed))

bigramUncrossedFrequency = {k: bigramUncrossedAmount[k] / len(bigramUncrossed)
for k in bigramUncrossedAmount}


#--- H1 ---


preH1 = []

for f in frequency.values():

    preH1.append(-f*math.log(f,2))

preH1 = sorted(preH1, reverse = 1)

H1 = sum(preH1)

print("\n/// Ентропія:", H1, "///")


#--- H2 ---


def specific_entropy(bigram,bigramFrequency):

    preH2 = []

    for f in bigramFrequency.values():

        preH2.append(-f*math.log(f,2))

    preH2 = sorted(preH2)

    H2 = sum(preH2)/2

    return H2


H2_Crossed = specific_entropy(bigram,bigramFrequency)

H2_Uncrossed = specific_entropy(bigramUncrossed,bigramUncrossedFrequency)


print("/// Питома ентропія на символ пересічної біграми:", H2_Crossed, "///")

print("/// Питома ентропія на символ непересічної біграми:", H2_Uncrossed,
"///")



alphabetFiltred = sorted(alphabetFiltred)

alphabetFiltred.insert(-27, "ё")
```

```
alphabetFiltred.pop()

df3 = pd.DataFrame(index = alphabetFiltred, columns=alphabetFiltred)

bigramList = []

for i in alphabetFiltred:
    for j in alphabetFiltred:
        bigramList.append(i+j)

n = 0
for i in range(0,len(alphabetFiltred)):
    df3[alphabetFiltred[i]] = bigramList[n:len(alphabetFiltred)+n]
    n = len(alphabetFiltred)+n

df3 = df3.T

for i in list(bigramFrequency.keys()):
    x,y = np.where(df3 == i)
    df3.iloc[x,y] = bigramFrequency[i]

for i in bigramList:
    x,y = np.where(df3 == i)
    df3.iloc[x,y] = 0

if " " in df3.index:
    df3 = df3.rename(index={" ": "пробіл"}, columns={" ": "пробіл"})

print("\n/// Таблиця частот пересічних біграм ///")
print(df3)


df4 = pd.DataFrame(index = alphabetFiltred, columns=alphabetFiltred)
```

```
n = 0
for i in range(0,len(alphabetFiltred)):
    df4[alphabetFiltred[i]] = bigramList[n:len(alphabetFiltred)+n]
    n = len(alphabetFiltred)+n

df4 = df4.T

for i in list(bigramUncrossedFrequency.keys()):
    x,y = np.where(df4 == i)
    df4.iloc[x,y] = bigramUncrossedFrequency[i]

for i in bigramList:
    x,y = np.where(df4 == i)
    df4.iloc[x,y] = 0

if " " in df4.index:
    df4 = df4.rename(index={" ": "пробіл"}, columns={" ": "пробіл"})

print("\n/// Таблиця частот непересічних біграм ///")
print(df4)

if "пробіл" in df.index:
    df.to_excel("Amount_with_space.xlsx")
    df2.to_excel("Frequency_with_space.xlsx")
    df3.to_excel("Bigram_Crossed_Frequency_with_space.xlsx")
    df4.to_excel("Bigram_Uncrossed_Frequency_with_space.xlsx")
else:
    df.to_excel("Amount_without_space.xlsx")
    df2.to_excel("Frequency_without_space.xlsx")
    df3.to_excel("Bigram_Crossed_Frequency_without_space.xlsx")
    df4.to_excel("Bigram_Uncrossed_Frequency_without_space.xlsx")
```

Приклад обраного очищеного тесту

```
/// Тест з пробілами чи без? ///

+ з пробілами, - без пробілів: +
первая книга моисеева бытие глава  в начале сотворил бог небо и землю  земля же была безвидна и пуста и
тьма над бездною и дух божий носился над водою  и сказал бог да будет свет и стал свет  и увидел бог
свет что он хорош и отделил бог свет от тьмы  и назвал бог свет днем а тьму ночью и был вечер и было
утро день один  и сказал бог да будет твердь посреди воды и да отделяет она воду от воды и стало так  и
создал бог твердь и отделил воду которая под твердью от воды которая над твердью и стало так  и назвал
бог твердь небом и увидел бог что это хорошо и был вечер и было утро день второй  и сказал бог да
соберется вода которая под небом в одно место и да явится суша и стало так и собралась вода под небом в
свои места и явилась суша  и назвал бог сушу землею а собрание вод назвал морями и увидел бог что это
хорошо  и сказал бог да произрастит земля зелень траву сеющую семя по роду и по подобию ее и дерево
плодовитое приносящее по роду своему плод в котором семя его на земле и стало так  и произвела земля
зелень траву сеющую семя по роду и по подобию ее и дерево плодовитое приносящее плод в котором семя его
по роду его на земле и увидел бог что это хорошо  и был вечер и было утро день третий  и сказал бог да
будут светила на тверди небесной для освещения земли и для отделения дня от ночи и для знамений и времен
и дней и годов  и да будут они светильниками на тверди небесной чтобы светить на землю и стало так  и
создал бог два светила великие светило большее для управления днем и светило меньшее для управления
ночью и звезды  и поставил их бог на тверди небесной чтобы светить на землю и управлять днем и ночью и
отделять свет от тьмы и увидел бог что это хорошо  и был вечер и было утро день четвёртый  и сказал бог
да произведет вода пресмыкающихся душу живую и птицы да полетят над землею по тверди небесной и стало
так  и сотворил бог рыб больших и всякую душу животных пресмыкающихся которых произвела вода по роду их
и всякую птицу пернатую по роду ее и увидел бог что это хорошо  и благословил их бог говоря плодитесь и
размножайтесь и наполняйте воды в морях и птицы да размножаются на земле  и был вечер и было утро день
пятый  и сказал бог да произведет земля душу живую по роду ее скотов и гадов и зверей земных по роду их
и стало так  и создал бог зверей земных по роду их и скот по роду его и всех гадов земных по роду их и
```

```
/// Тест з пробілами чи без? ///

+ з пробілами, - без пробілів: -
первыакнигамоисеевабытиеглававначалесотворилбогнебоиземлюземляжебылабезвиднаипустаитьманадбездноюидухбож
ийносилсянадводоюисказалбогдабудетсветисталсветиувиделбогсветчтоонхорошиотделилбогсветоттьмыиназвалбогсв
етднематьмуночьюибылвечерибылоутроденьодинисказалбогдабудеттвердьпосредиводыидаотделяетонаводуотводыиста
```

Кількість літер в тексті

```
/// Словник, де ключ - це літера, а значення - кількість цієї літери в тексті ///
{' ': 195525, 'а': 67456, 'б': 14938, 'в': 43690, 'г': 20000, 'д': 31609, 'е': 76046, 'ж': 8418, 'з':
14656, 'и': 77913, 'й': 9064, 'к': 23185, 'л': 40989, 'м': 27686, 'н': 46995, 'о': 95976, 'п': 22433,
'р': 33898, 'с': 49797, 'т': 46748, 'у': 21972, 'ф': 2460, 'х': 9177, 'ц': 4113, 'ч': 7842, 'ш': 6941,
'щ': 2407, 'ъ': 122, 'ы': 16267, 'ь': 13747, 'э': 872, 'ю': 5848, 'я': 16373, 'ё': 48}
```

Частота появи кожної літери в тексті

```
/// Частота появи кожної літери алфавіту в тексті ///
{' ': 0.18529469461557926, 'а': 0.06392655118265446, 'б': 0.014156410424076322, 'в': 0.0414040414665882,
'г': 0.01895355526051188, 'д': 0.029955146411475998, 'е': 0.07206710316704432, 'ж': 0.00797755140914945,
'з': 0.013889165294903104, 'и': 0.0738364175506131, 'й': 0.008589751244063984, 'к':
0.021971908935748394, 'л': 0.03884436382865607, 'м': 0.026237406547126595, 'н': 0.04453611647338779,
'о': 0.09095432098414441, 'п': 0.02125925525795315, 'р': 0.03212438081104158, 'с': 0.0471915095653855,
'т': 0.044302040065920464, 'у': 0.02082237580919835, 'ф': 0.002331287297042961, 'х':
0.008696838831285875, 'ц': 0.0038977986393242676, 'ч': 0.007431689017646708, 'ш': 0.0065778313531606476,
'щ': 0.0022810603756026047, 'ъ': 0.00011561668708912245, 'ы': 0.015415874171137336, 'ь':
0.013027726208312839, 'э': 0.0008263750093583179, 'ю': 0.005542019558173673 5, 'я': 0.01551632801401805,
'ё': 4.548853262522851e-05}
```

Перше значення - H_1, друге значення - H_2 для біграми з перетином літер, третє значення - H_2 для біграми без перетину літер

```
/// Ентропія: 4.300186751612514 ///
/// Питома ентропія на символ пересічної біграми: 3.905232070072632 ///
/// Питома ентропія на символ непересічної біграми: 3.9052292387205223 ///
```

Таблиця частот біграм з перетином літер

```
/// Таблиця частот пересічних біграм ///
             пробіл            а  ...            ю            я
пробіл   0.00900769   0.00271415  ...  0.000111826    0.0016319
а         0.0169909   0.00101591  ...  0.000490898  0.000489002
б        0.000393287  0.000422665  ...            0  0.000708864
в         0.00717488   0.00478199  ...            0  0.000947679
г         0.00105098   0.00112205  ...            0            0
д         0.00235214   0.00564437  ...  5.68607e-06  0.000228391
е          0.0154974  6.82329e-05  ...  0.000678538  0.000326949
ё        4.35932e-05            0  ...            0            0
ж        6.15991e-05  0.000623573  ...            0            0
з         0.00184134     0.004655  ...  5.68607e-06  0.000460572
и          0.0300101   0.00063684  ...  0.000316525    0.0013059
й         0.00680718            0  ...            0  3.79071e-06
к         0.00394613    0.0054899  ...            0            0
л         0.00839833   0.00508051  ...    0.0010396   0.00191147
м         0.00779466   0.00155135  ...            0  0.000532595
н         0.00418969    0.0104586  ...  0.000169634    0.0013457
о            0.015795  0.000238815  ...   0.00139404  0.000437828
п        3.50641e-05  0.000817847  ...            0   0.00038381
р        0.000595142   0.00714076  ...  0.000198065   0.00077141
с         0.00269804   0.00194274  ...  0.000141204   0.00286388
т         0.00720899   0.00414136  ...  5.68607e-06  0.000643474
у         0.00809602  2.17966e-05  ...  0.000652003  1.51629e-05
ф        0.000204699  0.000815004  ...            0  1.80059e-05
х          0.0055534  0.000519328  ...  9.47679e-07            0
ц        0.000434985   0.00206499  ...            0            0
ч        0.000158262  0.000863335  ...            0            0
ш        0.000198065  0.000803632  ...            0            0
щ        6.63375e-06  0.000286199  ...            0            0
ъ                  0            0  ...            0  8.15004e-05
ы         0.00566807            0  ...            0            0
```

Таблиця частот біграм з перетином літер

```
/// Таблиця частот непересічних біграм ///
             пробіл            а  ...            ю            я
пробіл   0.00911667   0.00275964  ...  0.000123198   0.00165844
а         0.0170127   0.00104813  ...   0.00047763    0.0005307
б       0.000394234  0.000422665  ...            0   0.00071455
в        0.00718151   0.00475735  ...            0  0.000949574
г        0.00101212   0.00113911  ...            0            0
д        0.00241658   0.00560836  ...  5.68607e-06  0.000221757
е         0.0156196  6.25468e-05  ...  0.000669061  0.000307048
ё       5.11746e-05            0  ...            0            0
ж       5.87561e-05  0.000615991  ...            0            0
з        0.00183471    0.0045318  ...  5.68607e-06  0.000426455
и         0.0301021  0.000657689  ...   0.00034685   0.00126231
й        0.00669251            0  ...            0  5.68607e-06
к        0.00389875    0.0053904  ...            0            0
л        0.00838885   0.00516295  ...   0.00107088    0.0019598
м        0.00776149   0.00154472  ...            0  0.000547758
н        0.00420959    0.0103676  ...  0.000153524   0.00131348
о         0.0159058  0.000227443  ...    0.0013874  0.000434037
п       4.35932e-05  0.000837748  ...            0  0.000344955
р       0.000570503   0.00712275  ...   0.00018764  0.000765724
с        0.00260612   0.00188967  ...   0.00015921   0.00279565
т        0.00715497   0.00400489  ...  5.68607e-06  0.000663375
у        0.00799841  1.89536e-05  ...   0.00063684  7.58143e-06
ф       0.000197117   0.00077141  ...            0  2.08489e-05
х        0.00549464  0.000507956  ...  1.89536e-06            0
ц       0.000407502   0.00213417  ...            0            0
ч       0.000155419  0.000885132  ...            0            0
ш       0.000200908  0.000805527  ...            0            0
щ       7.58143e-06  0.000303257  ...            0            0
ъ                 0            0  ...            0  8.52911e-05
ы        0.00564248            0  ...            0            0
ь        0.00965305  1.89536e-06  ...  0.000233129  0.000388548
```

Результати виконання роботи

Таблиця з назвами файлів, які містять значення частот

	Текст з пробілами	Текст без пробілів
Букви	"Frequency_with_space.xlsx"	"Frequency_without_space.xlsx"
Біграми з перетином	"Bigram_Crossed_Frequency_with_space.xlsx"	"Bigram_Crossed_Frequency_without_space.xlsx"
Біграми без перетину	"Bigram_Uncrossed_Frequency_with_space.xlsx"	"Bigram_Uncrossed_Frequency_without_space.xlsx"

Таблиця зі значеннями ентропії

Позначення	Текст з пробілами	Текст без пробілів
H_1	4.300186752	4.429408817
H_2 з перетином	3.905232070	4.115231470
H_2 без перетину	3.905229239	4.115063150

Таблиця з оцінкою надлишковості R мови

$$R = 1 - \frac{H_n}{H_0}$$

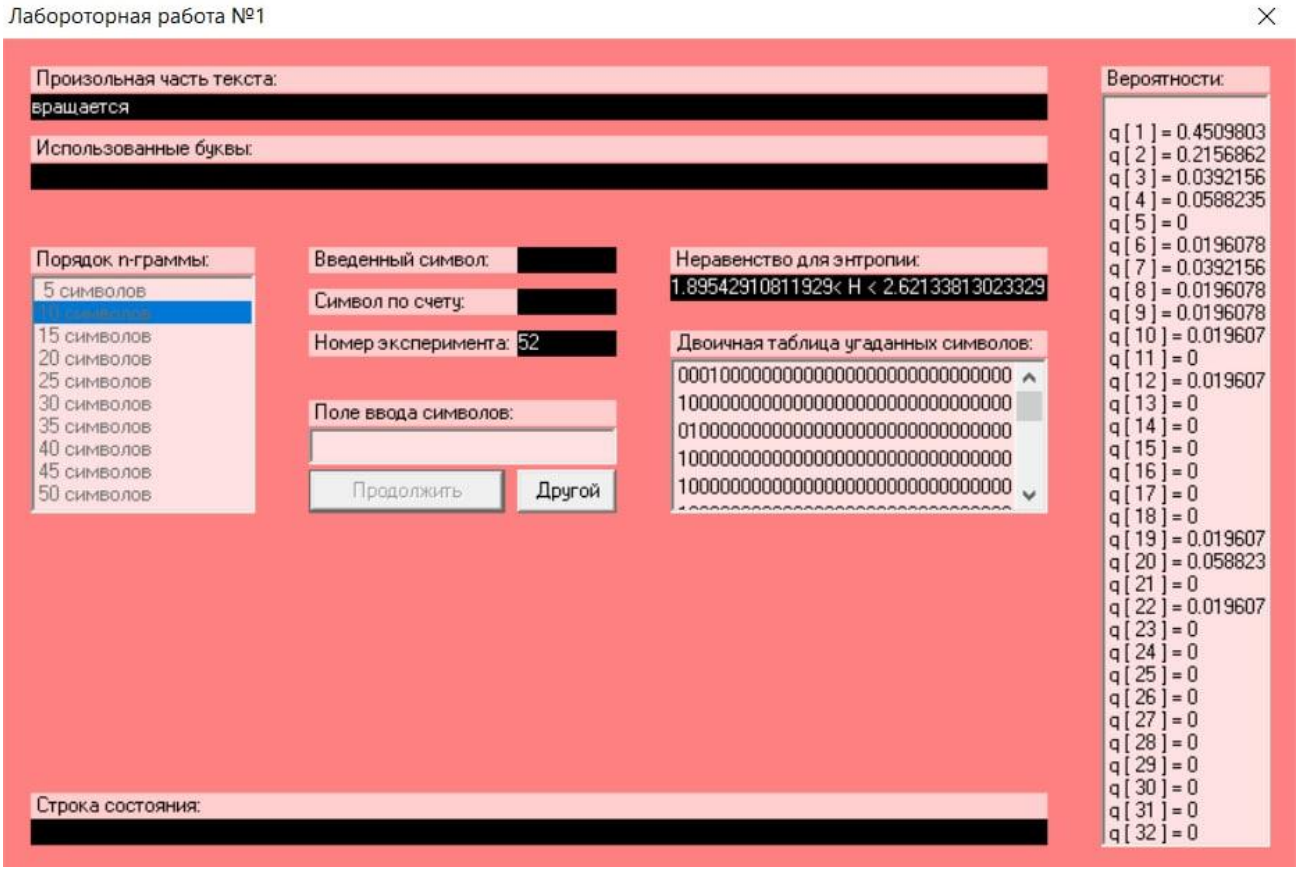
	Текст з пробілами	Текст без пробілів
H_1	0.154748273	0.121914602
H_2 з перетином	0.232381210	0.184197077
H_2 без перетину	0.232381767	0.184230444

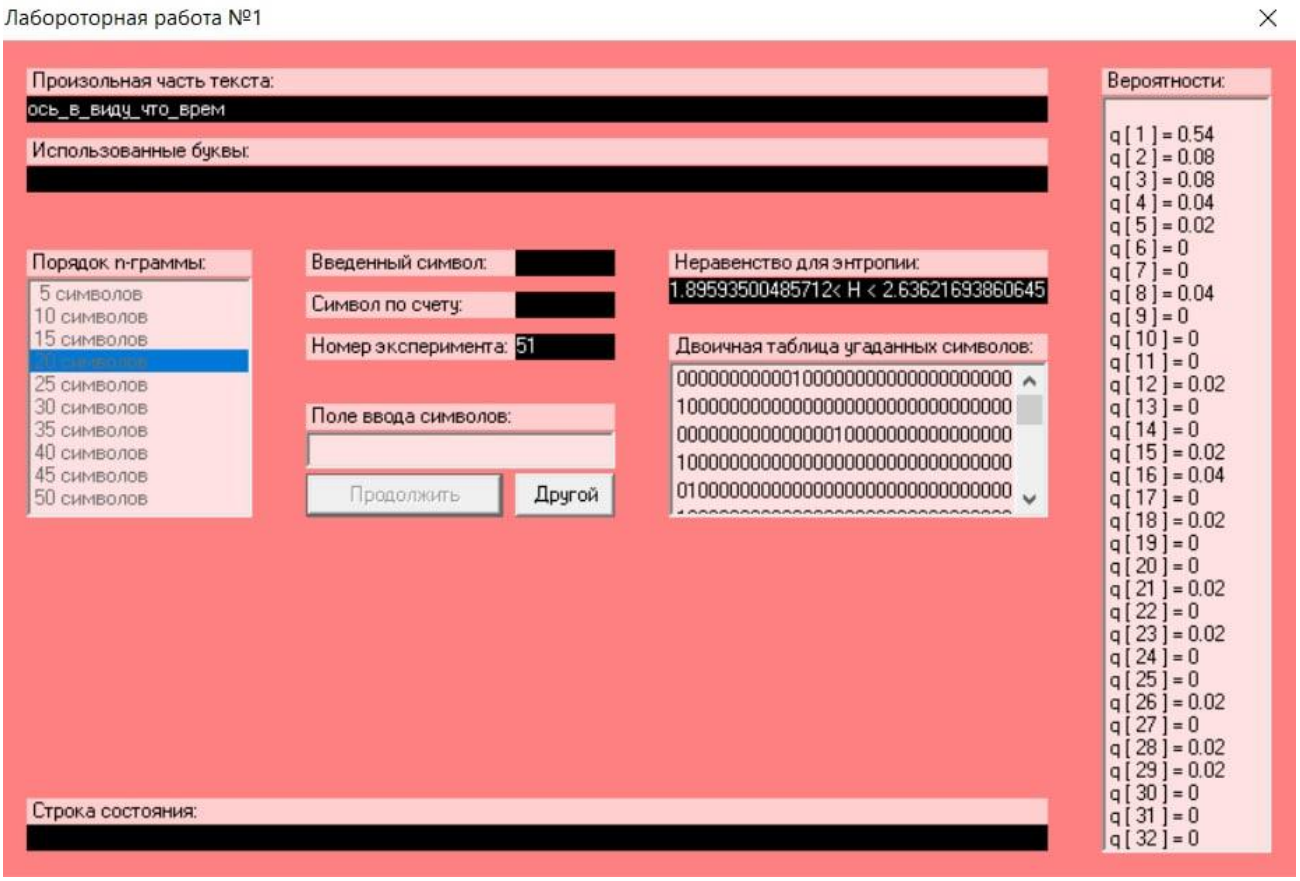
$1.895429108 < H^{(10)} < 2.621338130$ $0.6209141784 > R^{(10)} > 0.475732374$

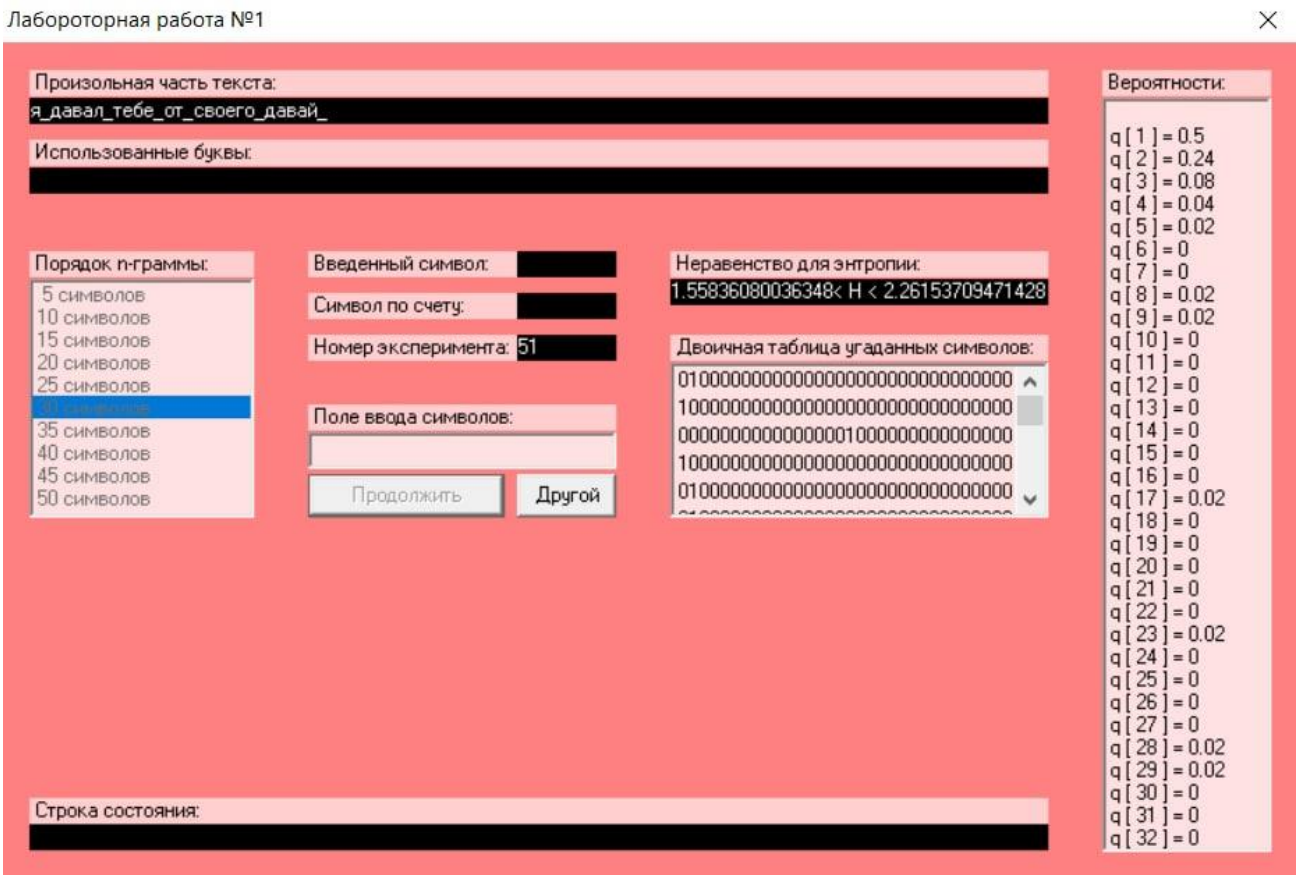
$1.895935004 < H^{(20)} < 2.636216938$ $0.6208129992 > R^{(20)} > 0.4727566124$

$1.558360800 < H^{(30)} < 2.261537094$ $0.68832784 > R^{(30)} > 0.5476925812$

CoolPinkProgram




Лабороторная работа №1
Произольная часть текста:
ось_в_виду_что_врем
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:
1.89593500485712< H < 2.63621693860645
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0.54
q [2] = 0.08
q [3] = 0.08
q [4] = 0.04
q [5] = 0.02
q [6] = 0
q [7] = 0
q [8] = 0.04
q [9] = 0
q [10] = 0
q [11] = 0
q [12] = 0.02
q [13] = 0
q [14] = 0
q [15] = 0.02
q [16] = 0.04
q [17] = 0
q [18] = 0.02
q [19] = 0
q [20] = 0
q [21] = 0.02
q [22] = 0
q [23] = 0.02
q [24] = 0
q [25] = 0
q [26] = 0.02
q [27] = 0
q [28] = 0.02
q [29] = 0.02
q [30] = 0
q [31] = 0
q [32] = 0


Лабороторная работа №1
Произольная часть текста:
я_давал_тебе_от_своего_давай_
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:
1.55836080036348< H < 2.26153709471428
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0.5
q [2] = 0.24
q [3] = 0.08
q [4] = 0.04
q [5] = 0.02
q [6] = 0
q [7] = 0
q [8] = 0.02
q [9] = 0.02
q [10] = 0
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0
q [17] = 0.02
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0.02
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0.02
q [29] = 0.02
q [30] = 0
q [31] = 0
q [32] = 0

$1.895429108 < H^{(10)} < 2.621338130$

$1.895935004 < H^{(20)} < 2.636216938$

$1.558360800 < H^{(30)} < 2.261537094$

Висновок

У ході виконання лабораторної роботи, ми ближче познайомились із поняттям ентропії та навчились обчислювати її на практиці. Зазначимо, що ентроія зменшується за наявності пробілу в алфавіті, адже пробіл має найбільшу частоту появи у тексті. Також ми оцінили надлишковість російської мови, вона становить 0,12 - 0,23. У ході роботі з даною нам програмою CoolPinkProgram ми обчислили нерівності для ентропій кожного з експериментів ($H^{(10)}$, $H^{(20)}$, $H^{(30)}$).