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

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

з предмету «Криптографія»

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконав:
Студент 3 курсу,
ФТІ, групи ФБ-05
Савченко Ярослав

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

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

Хід роботи:

Напишемо наш прототип програми для виконання завдання 1

Файл lab1_1.py містить в собі 3 функції для знаходження частоти літер, біграм та для знаходження ентропії. Пропишемо якийсь довільний текст для перевірки

```
    return entropy

text = "Этот текст содержит в себе какие-то странные штуки, не смотрите сюда и не испугаетесь. А ещё я люблю котиков <3"
letter_freqs = frequency_letter(text)
bigram_freqs = frequency_bigram(text)
```

Перевіримо

```
C:\CryptLab\crypto-22-23\cp1\savchenko_fb-05_cp1>python lab1.py
Частота букавок: {'э': 0.011494252873563218, 'т': 0.13793103448275862, 'о': 0.06896551724137931, 'е': 0.13793103448275862, 'к': 0.06896551724137931, 'с': 0.09195402298850575, 'д': 0.022988505747126436, 'р': 0.034482758620689655, 'ж': 0.011494252873563218, 'и': 0.08045977011494253, 'в': 0.022988505747126436, 'б': 0.022988505747126436, 'а': 0.05747126436781609, 'н': 0.04597701149425287, 'ы': 0.011494252873563218, 'ш': 0.011494252873563218, 'у': 0.022988505747126436, 'м': 0.011494252873563218, 'ю': 0.034482758620689655, 'п': 0.011494252873563218, 'г': 0.011494252873563218, 'ь': 0.011494252873563218, 'щ': 0.011494252873563218, 'ё': 0.011494252873563218, 'я': 0.011494252873563218, 'л': 0.022988505747126436}
Частота биграмм: {'эт': 0.014925373134328358, 'то': 0.029850746268656716, 'от': 0.04477611940298507, 'те': 0.04477611940298507, 'ек': 0.014925373134328358, 'кс': 0.014925373134328358, 'ст': 0.029850746268656716, 'со': 0.014925373134328358, 'од': 0.014925373134328358, 'де': 0.014925373134328358, 'ер': 0.014925373134328358, 'рж': 0.014925373134328358, 'жи': 0.014925373134328358, 'ит': 0.029850746268656716, 'се': 0.014925373134328358, 'еб': 0.014925373134328358, 'бе': 0.014925373134328358, 'ка': 0.014925373134328358, 'ак': 0.014925373134328358, 'ки': 0.029850746268656716, 'ие': 0.014925373134328358, 'тр': 0.029850746268656716, 'ра': 0.014925373134328358, 'ан': 0.014925373134328358, 'нн': 0.014925373134328358, 'ны': 0.014925373134328358, 'ые': 0.014925373134328358, 'шт': 0.014925373134328358, 'ту': 0.014925373134328358, 'ук': 0.014925373134328358, 'не': 0.029850746268656716, 'см': 0.014925373134328358, 'мо': 0.014925373134328358, 'ри': 0.014925373134328358, 'сю': 0.014925373134328358, 'юд': 0.014925373134328358, 'да': 0.014925373134328358, 'ис': 0.014925373134328358, 'сп': 0.014925373134328358, 'пу': 0.014925373134328358, 'уг': 0.014925373134328358, 'га': 0.014925373134328358, 'ае': 0.014925373134328358, 'ет': 0.014925373134328358, 'ес': 0.014925373134328358, 'сь': 0.014925373134328358, 'ещ': 0.014925373134328358, 'щё': 0.014925373134328358, 'лю': 0.029850746268656716, 'юб': 0.014925373134328358, 'бл': 0.014925373134328358, 'ко': 0.029850746268656716, 'ти': 0.014925373134328358, 'ик': 0.014925373134328358, 'ов': 0.014925373134328358}
Энтропия букавок: 4.146052117008171
Энтропия биграм: 5.685346279945427
Тыкни лапкой что бы закрыть консоль...
```

Чудово. Задачу виконав, але трохи модифікуємо програму. Додамо можливість записати результати частот у файл, для легшої роботи з даними. А також додамо можливість брати текст з файлу.

Модифікована версія знаходиться в файлі lab1_2.py

Тестуємо. В текстовий файл я додав видання мого олюбленого письменника Лавкрафта що розмір був більше 1 МБ

```
45      text = f.read()

Exception has occurred: UnicodeDecodeError ×
'charmap' codec can't decode byte 0x98 in position 1413: character maps to <undefined>
  File "C:\CryptLab\crypto-22-23\cp1\savchenko_fb-05_cp1\lab1_2.py", line 45, in <module>
    text = f.read()
UnicodeDecodeError: 'charmap' codec can't decode byte 0x98 in position 1413: character maps to <undefined>

46
```

Перша проблема. Але вирішується простим ігноруванням символів які не можуть розпізнатися Unicode. Адже нам такі не потрібні.

```
-' 'c:\CryptLab\crypto-22-23\cp1\savchenko_fb-05_cp1\lab1_2.py'
Частота букавок записана в файл letter_frequencies.txt
Частота биграмм записана в файл bigram_frequencies.txt
Энтропия букавок: 4.470956362742454
Энтропия биграмм: 7.978130755582727
Тыкни лапкой что бы закрыть консоль...
PS C:\CryptLab\crypto-22-23\cp1\savchenko_fb-05_cp1>
```

Чудово, програма виконала свою задачу, подивимося на файли

letter_frequencies.txt – Блокнот

Файл Правка Формат Вид Справка

```
п: 0.027591261835135522
р: 0.04472604447668488
и: 0.07480787976333507
ж: 0.009425792850180218
з: 0.01845261034162265
н: 0.06955168166602438
э: 0.0034744104156748954
т: 0.056442923099010875
о: 0.11156499595735195
с: 0.05518705748116579
а: 0.07247296715254006
е: 0.08595954329411587
л: 0.04779392318213074
ь: 0.017429480349707947
у: 0.025865391154534945
б: 0.01607235962188017
к: 0.03125609230838982
в: 0.043940183921595294
д: 0.02938211713856082
й: 0.011967749491835362
г: 0.017267774427795285
м: 0.03611482631726098
ч: 0.014629096487052191
я: 0.02126961817755915
ш: 0.007715034872562132
ы: 0.021673127347191682
х: 0.011789419596641957
ц: 0.0029983602717263995
щ: 0.004644133626519771
ё: 1.5112702982491934e-06
ф: 0.002194364473057829
ю: 0.005859194946312122
і: 2.26690544737379e-05
```

bigram_frequencies.txt – Блокнот

Файл Правка Формат Вид Справка

```
пр: 0.009258011954036563
ри: 0.005875525226560179
жи: 0.00205385036971152
из: 0.004249791016572893
зн: 0.0025502436755986705
ни: 0.012863783402748432
эт: 0.002683107311375157
то: 0.01607096394413084
от: 0.008433150215257543
пи: 0.0016478781492833667
ис: 0.00555812876331635
са: 0.00233434026746188
ат: 0.006183695048430641
те: 0.0073517878462989175
ел: 0.0074717341841526895
ль: 0.005727898964586305
не: 0.013670191858780718
оп: 0.002159034081367905
пу: 0.0011791647675163173
уб: 0.0008396243649764075
бл: 0.0014688813066400447
ли: 0.011778730377240459
ик: 0.0026259021348602808
ко: 0.011380139469911
ов: 0.009479451346997373
ва: 0.007305654639432082
ал: 0.011097804243885966
од: 0.0069808768630895595
дн: 0.002788291023031542
но: 0.016074654600680187
ой: 0.005761114873530426
кн: 0.0005849690630714751
иг: 0.0008562323194484683
```

Знову модифікуємо код, щоб відсортувати результат по спаданню.

Найбільш частими буквами та біграмами були:

о: 0.11156499595735195 е: 0.08595954329411587 и: 0.07480787976333507 а: 0.07247296715254006 н: 0.06955168166602438 т: 0.056442923099010875 с: 0.05518705748116579 л: 0.04779392318213074 р: 0.04472604447668488 в: 0.043940183921595294	ст: 0.017174470252385547 но: 0.016074654600680187 то: 0.01607096394413084 не: 0.013670191858780718 на: 0.013544709536102924 по: 0.01345059779409458 ни: 0.012863783402748432 ен: 0.01206475625981484 ра: 0.011872842119248803 ли: 0.011778730377240459

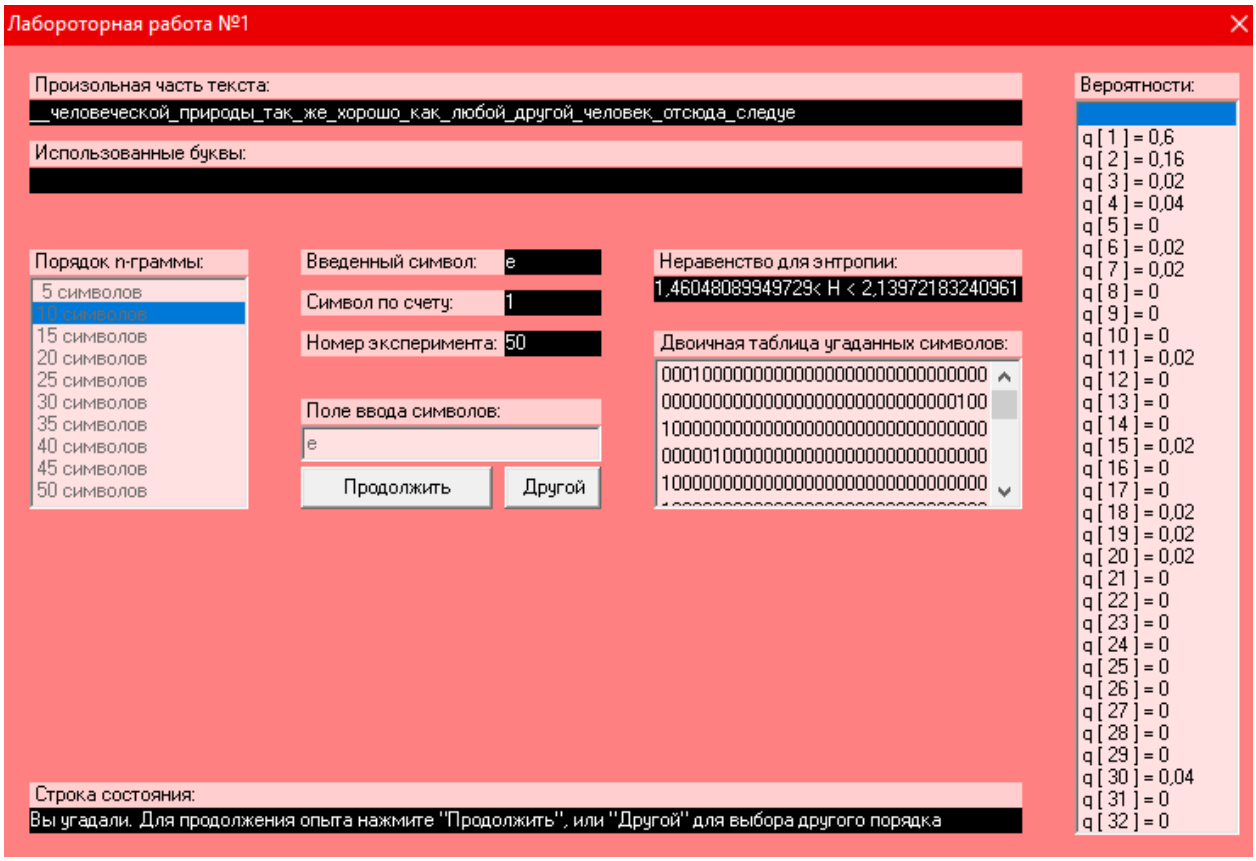
Ентропія:

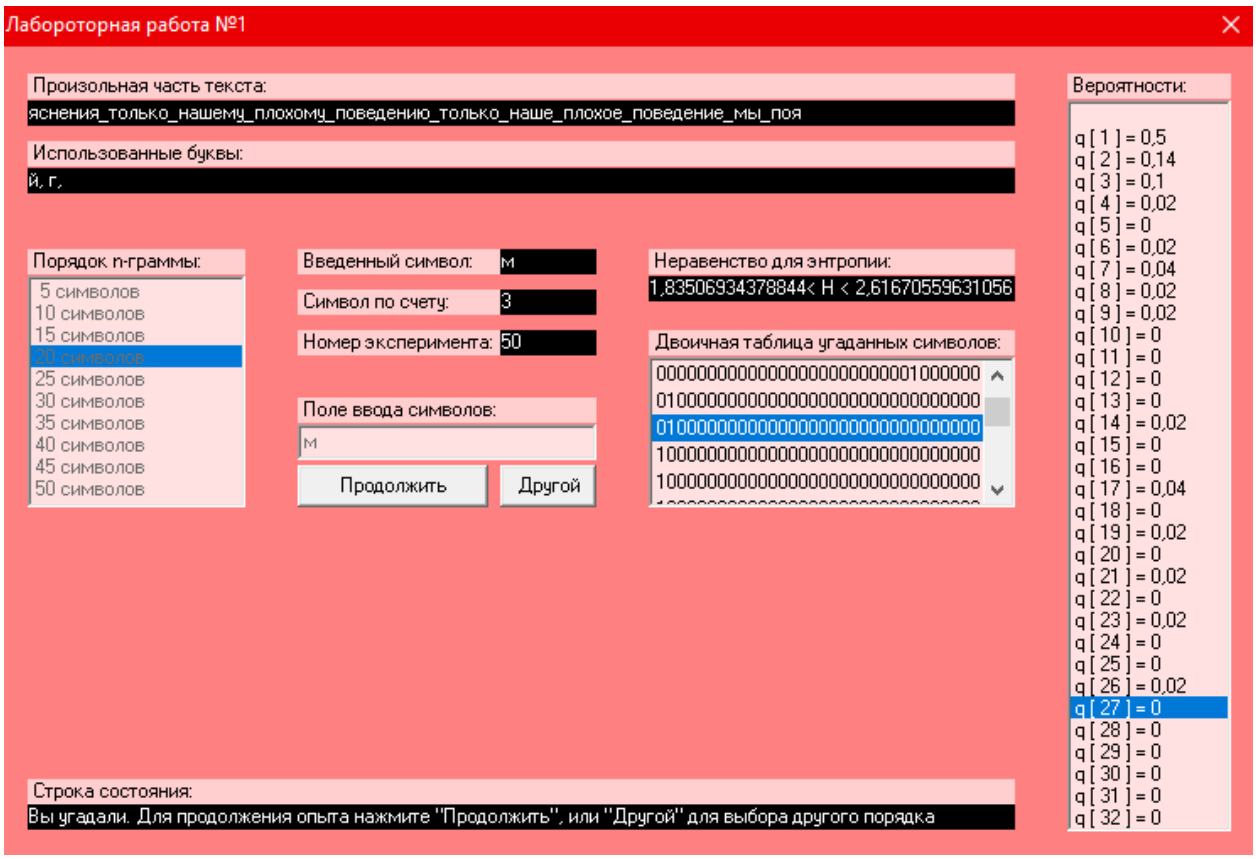
4.470956362742454	7.978130755582727

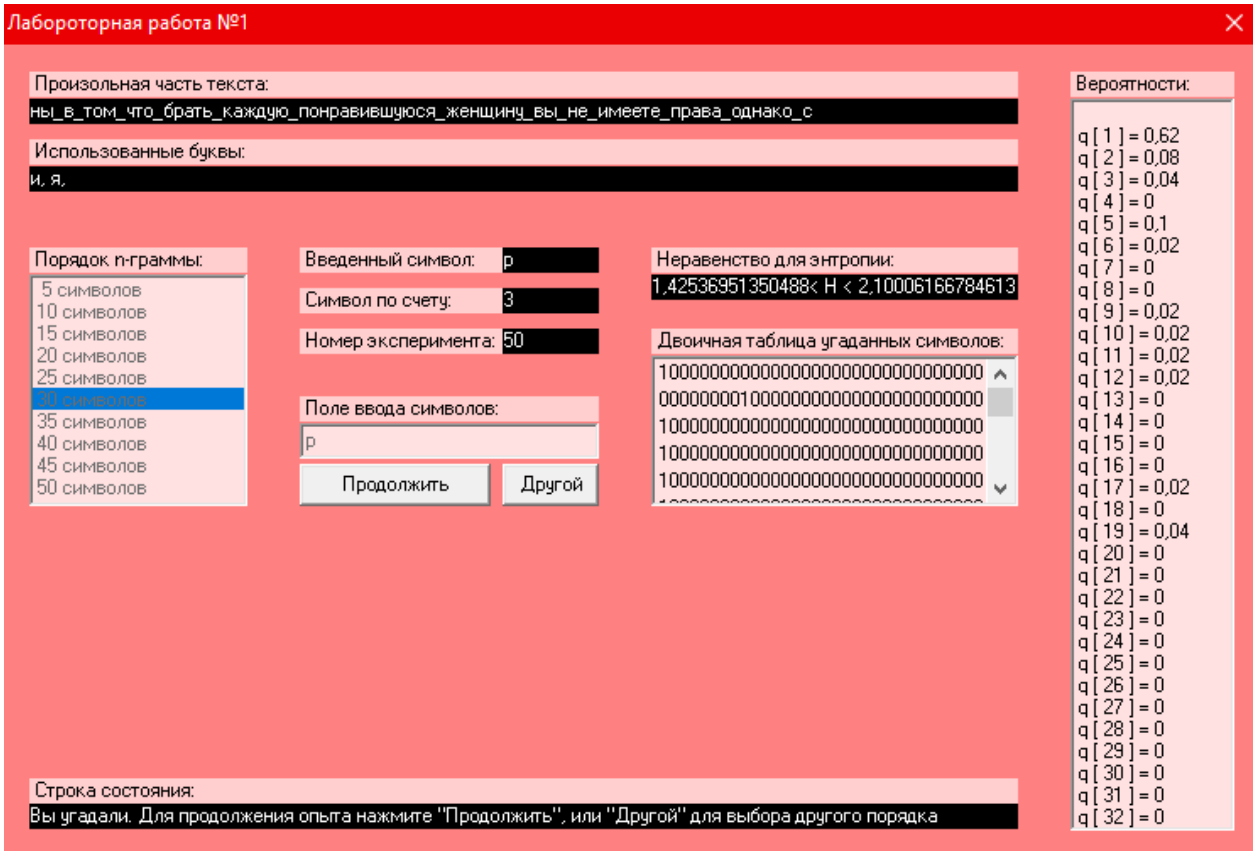
Надлишковість:

0,1058087275	0,2021869244

Перейдемо до другої задачі. Запустимо програму CoolPinkProgram

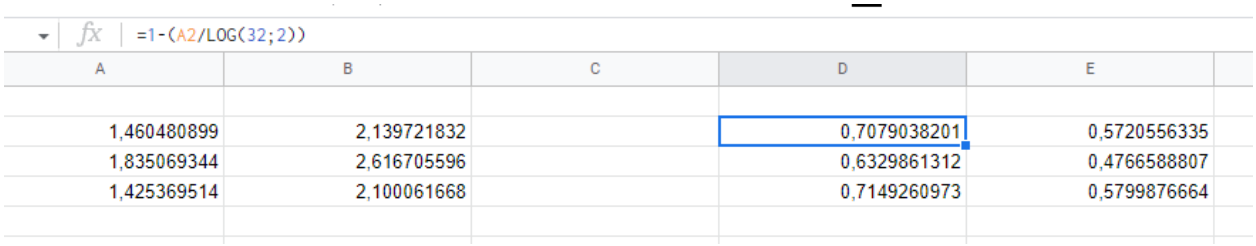






Тепер порахуємо оцінку надлишковості для кожного експерименту

Для цього я записав дані енропії в exel таблицю та функцією порахував надлишковість



Отже результат:

0,5720556335 < R < 0,7079038201

0,4766588807 < R < 0,6329861312

0,5799876664 < R < 0,7149260973

Висновки: на цій лабораторній роботі я отримав навички оцінки ентропії на символ джерела, навчився визначати наближене значення ентропії. А саме головне, це поборов помилку в якій використав не те кодування, ~~але таку безглузду проблему я не покажу~~ :)