мета: оцінити ентропію тексту

1. Знаходимо текст достатньої довжини

2. Запускаємо код для видалення знаків пунктуації та заміни великих літер на малі

```
import string

def process_text_file(input_file_path, output_file_path):
    with open(input_file_path, 'r', encoding='utf-8') as input_file,
open(output_file_path, 'w', encoding='utf-8') as output_file:
        text = input_file.read()
        cleaned_text = ''.join([char.lower() for char in text if char not in
string.punctuation])
        output_file.write(cleaned_text)

input_file_path = r'C:\Users\Sasha\Desktop\lab1.txt'
output_file_path = r'C:\Users\Sasha\Desktop\lab1_1.txt'

process_text_file(input_file_path, output_file_path)
```

3. Пишемо код для знаходження ентропії(lab1_1.py)

4. Отримуємо значення ентропії та надлишковості

```
Ентропія H1 для літер (з пробілами): 4.345264629675523
Ентропія H1 для літер (без пробілів): 4.446953916455404
Ентропія H2 для біграм (з пробілами, з перетинами): 3.942005067508832
Ентропія H2 для біграм (без пробілів, з перетинами): 4.1310073893262125
Ентропія H2 для біграм (з пробілами, без перетинів): 4.5943939320922595
Ентропія H2 для біграм (без пробілів, без перетинів): 4.1313161795224165
Надлишковість для літер (з пробілами): 0.14588769190742779
Надлишковість для літер (без пробілів): 0.1258994797173827
Надлишковість для біграм (з пробілами, з перетинами): 0.2185374547993929
Надлишковість для біграм (без пробілів, з перетинами): 0.1810696603833971
Надлишковість для біграм (з пробілами, без перетинів): 0.08920797555037685
Надлишковість для біграм (без пробілів, без перетинів): 0.18100844585715326
```

та таблиці частот літер та біграм

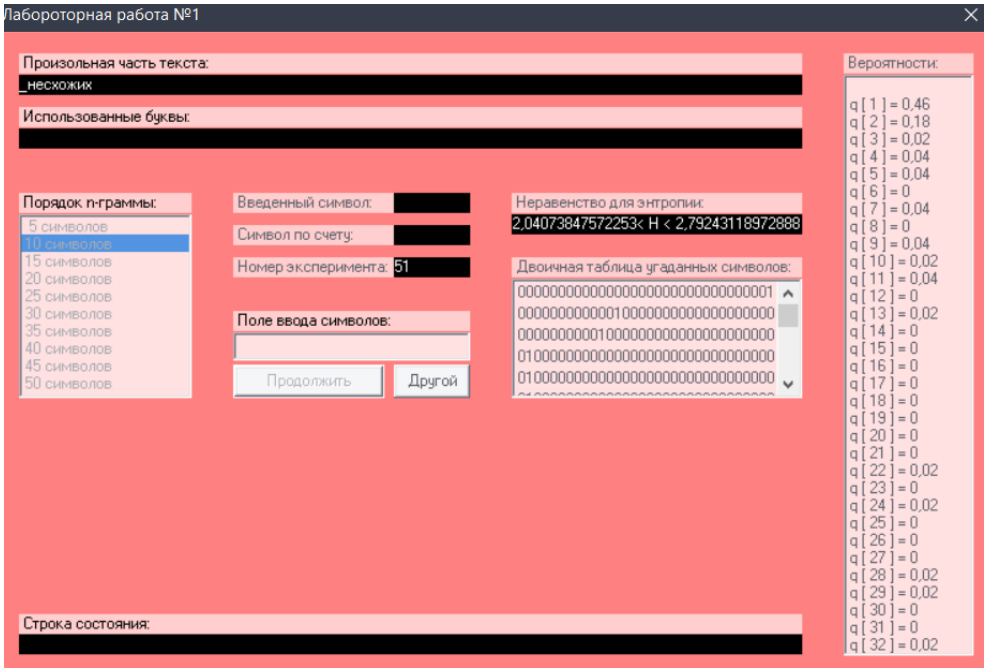
Character	Count	Frequency
	119982	17.157%
о	67040	9.587%
е	51059	7.301%
а	46476	6.646%
т	37971	5.430%
и	37713	5.393%
н	37019	5.294%
с	30400	4.347%
л	26929	3.851%
в	26530	3.794%
р	23832	3.408%
к	19204	2.746%
м	18291	2.616%
д	17793	2.544%
у	17439	2.494%

Частота біграм:

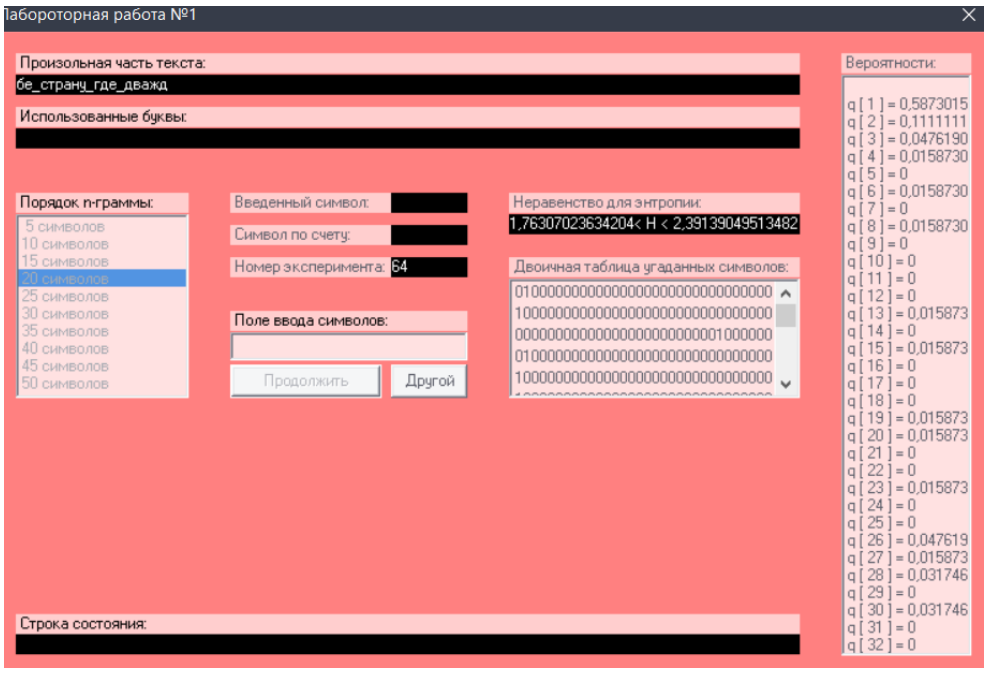
Character	Count	Frequency
о	17265	2.469%
е	13471	1.926%
и	12185	1.742%
а	11756	1.681%
в	11351	1.623%
н	11033	1.578%
с	10785	1.542%
то	10230	1.463%
п	10196	1.458%
о	9397	1.344%
ь	8655	1.238%
и	8101	1.158%
я	7663	1.096%

5. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30)

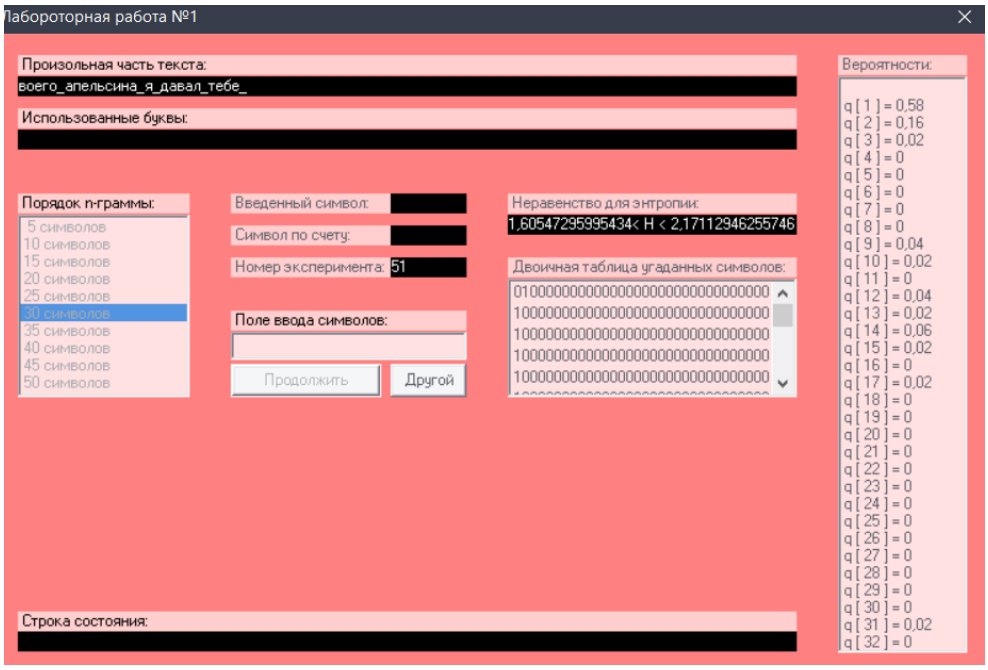
$$R = 1 - \frac{H}{\log_2 m}$$



0,15<R<0,38



0,44<R<0,59



0,55<R<0,66

Висновки:

У ході роботи ми засвоїли поняття ентропії на символ джерела та його надлишковості та практично обрахували значння ентропії для різних моделей джерела тексту. У мене виникли деякі труднощі з CoolPinkProgram, спочатку я просто рандомно нажимав літери на клавіатурі, а потім зрозумів що треба вгадувати на основі частини тексту, яка вже є)