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

```
Ентропія H1 для літер (з пробілами): 4.431447389078525
Ентропія H2 для біграм (з пробілами): 7.616771749622542
Ентропія H1 для літер (без пробілів): 4.42210334030845
Ентропія H2 для біграм (без пробілів): 7.7338024912633
```

та таблиці частот літер та біграм

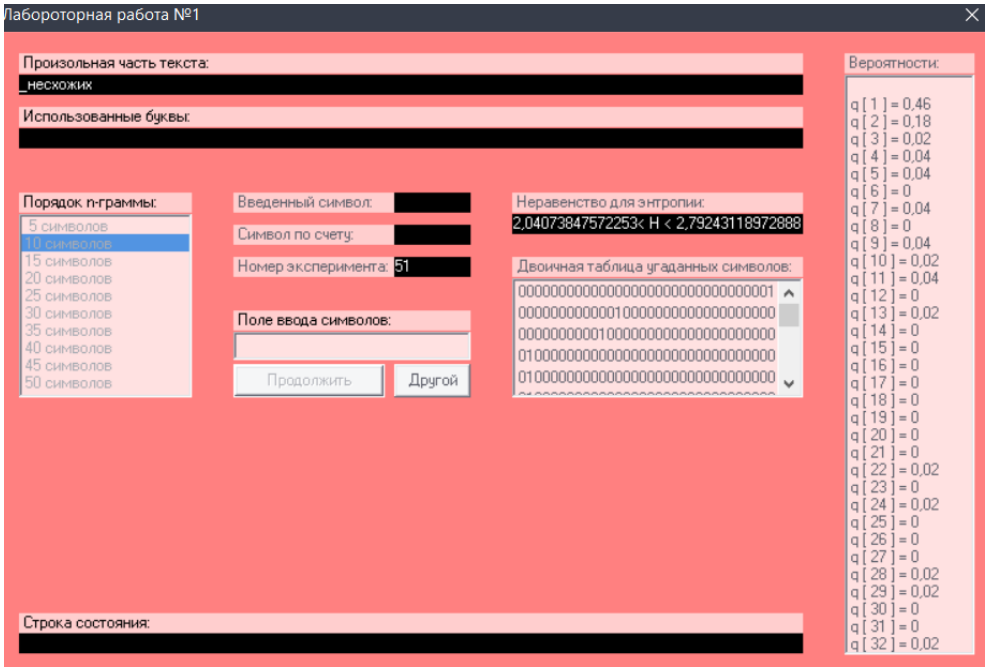
```
Частота літер (з пробілами):
+---------------+------------+-----------+
|   Character   |    Count   | Frequency |
+---------------+------------+-----------+
|               | 65672      | 11.290%   |
| о             | 56262      | 9.672%    |
| и             | 45514      | 7.825%    |
| а             | 44563      | 7.661%    |
| н             | 42837      | 7.364%    |
| е             | 41253      | 7.092%    |
| т             | 32665      | 5.616%    |
| с             | 28509      | 4.901%    |
| р             | 26318      | 4.525%    |
| в             | 24462      | 4.205%    |
| п             | 14609      | 2.512%    |
| л             | 14393      | 2.474%    |
```

Частота біграм:

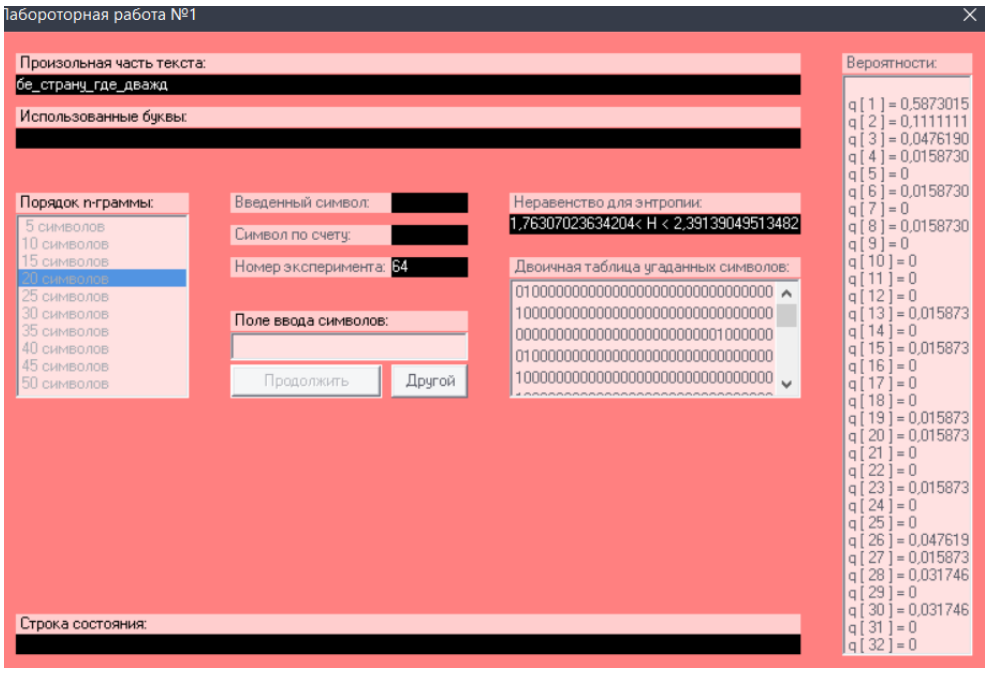
Character	Count	Frequency
ст	10761	1.850%
ни	10502	1.805%
п	9875	1.698%
ра	9385	1.613%
и	8947	1.538%
но	8793	1.512%
й	8371	1.439%
ов	8228	1.415%
ен	8212	1.412%
на	8001	1.376%

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

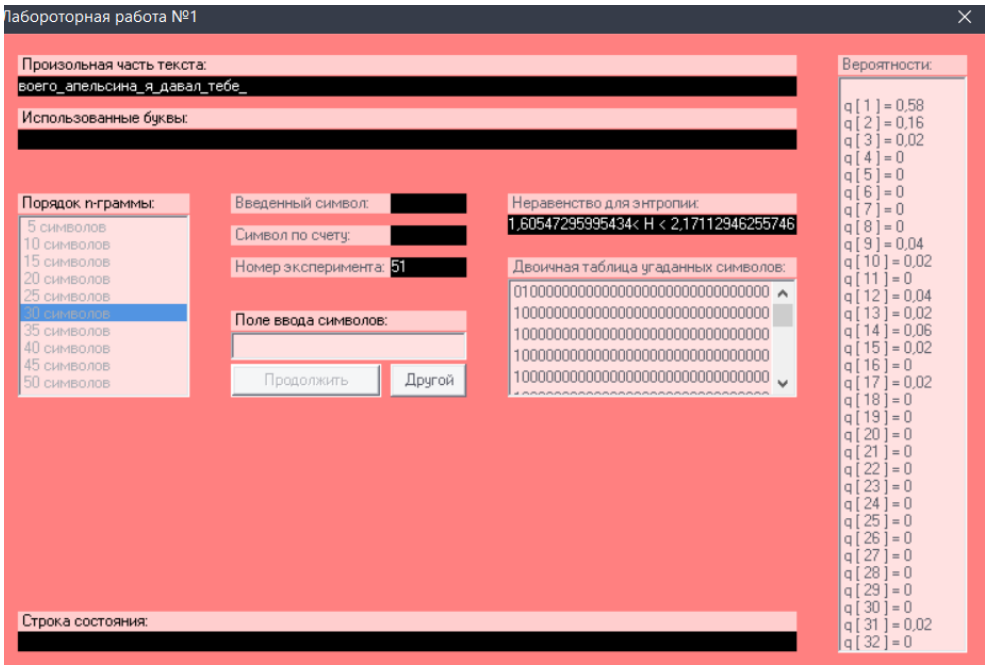
$$R = 1 - \frac{H}{\log_2 m}$$



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)