мета: оцінити ентропію тексту

- 1. Знаходимо текст достатньої довжини
- 2. Запускаємо код для видалення знаків пунктуації та заміни великих літер на малі

```
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 Н1 для літер (з пробілами): 4.345264629675523
Ентропія Н1 для літер (без пробілів): 4.446953916455404
Ентропія Н2 для біграм (з пробілами, з перетинами): 3.942005067508832
Ентропія Н2 для біграм (без пробілів, з перетинами): 4.1310073893262125
Ентропія Н2 для біграм (з пробілами, без перетинів): 2.220945705751801
Ентропія Н2 для біграм (без пробілів, без перетинів): 2.3156580897611914
```

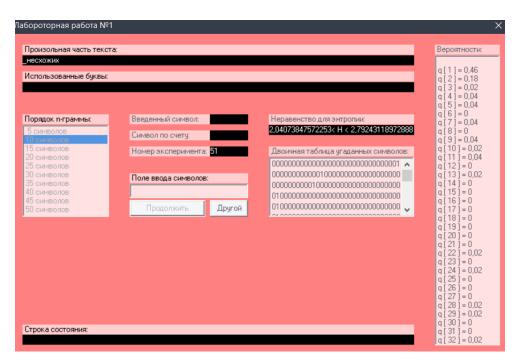
та таблиці частот літер та біграм

ŀ	Character		Count		Frequency	i I
+		+-		+-		+
1		Ī	119982	Τ	17.157%	1
o		Ī	67040	Ī	9.587%	L
e		Ī	51059	Ī	7.301%	L
a		Ī	46476	Ī	6.646%	L
т		Ī	37971	Ī	5.430%	L
и		Ĺ	37713	Ī	5.393%	L
н		Ī	37019	Ī	5.294%	L
c		Ī	30400	Ī	4.347%	L
л		Ī	26929	Ī	3.851%	L
в		Ī	26530	Ī	3.794%	L
p		Ĺ	23832	Ī	3.408%	L
к		Ī	19204	Ī	2.746%	L
м		Ī	18291	Ī	2.616%	T.
Ιд		Ī	17793	Ī	2.544%	T
Ιу		Ī	17439	Ī	2.494%	T

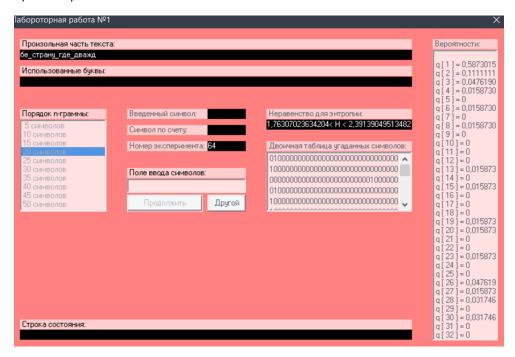
Частота біграм:				
		Count		Frequency
+	-+-	172/F	.+.	+
0 e	¦	17265 13471		2.469% 1.926%
и	i	12185		1.742%
a	Ī	11756	Ī	1.681%
в	1	11351	I	1.623%
н	1	11033		1.578%
C		10785		1.542%
то п	¦	10230 10196		1.463% 1.458%
 o	ı	9397		1.344%
ь	1	8655	Ī	1.238%
и	Ī	8101	Ī	1.158%
я	T	7663	T	1.096%

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

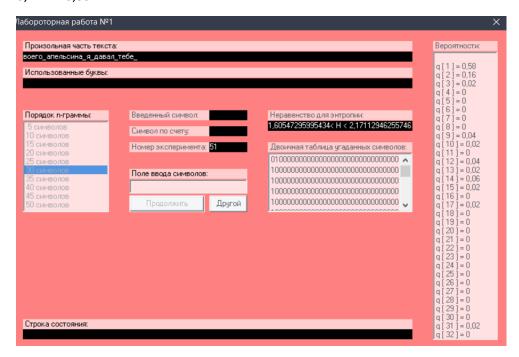
$$R = 1 - \frac{H}{\log_2 m}$$



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)