

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%  |
| п  | 14393 | 2.474%  |
```



Лабораторная работа №1

Произвольная часть текста:  
бе\_страну\_где\_дважд

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: 64

Неравенство для энтропии:  
 $1,76307023634204 < H < 2,39139049513482$

Двоичная таблица угаданных символов:

01000000000000000000000000000000
10000000000000000000000000000000
00000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

$q[1] = 0,5873015$
$q[2] = 0,1111111$
$q[3] = 0,0476190$
$q[4] = 0,0158730$
$q[5] = 0$
$q[6] = 0,0158730$
$q[7] = 0$
$q[8] = 0,0158730$
$q[9] = 0$
$q[10] = 0$
$q[11] = 0$
$q[12] = 0$
$q[13] = 0,015873$
$q[14] = 0$
$q[15] = 0,015873$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0,015873$
$q[20] = 0,015873$
$q[21] = 0$
$q[22] = 0$
$q[23] = 0,015873$
$q[24] = 0$
$q[25] = 0$
$q[26] = 0,047619$
$q[27] = 0,015873$
$q[28] = 0,031746$
$q[29] = 0$
$q[30] = 0,031746$
$q[31] = 0$
$q[32] = 0$

Строка состояния:

$0,44 < R < 0,59$

Лабораторная работа №1

Произвольная часть текста:  
воего\_апельсина\_я\_давал\_тебе\_

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

Неравенство для энтропии:  
 $1,60547295995434 < H < 2,17112946255746$

Двоичная таблица угаданных символов:

01000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Вероятности:

$q[1] = 0,58$
$q[2] = 0,16$
$q[3] = 0,02$
$q[4] = 0$
$q[5] = 0$
$q[6] = 0$
$q[7] = 0$
$q[8] = 0$
$q[9] = 0,04$
$q[10] = 0,02$
$q[11] = 0$
$q[12] = 0,04$
$q[13] = 0,02$
$q[14] = 0,06$
$q[15] = 0,02$
$q[16] = 0$
$q[17] = 0,02$
$q[18] = 0$
$q[19] = 0$
$q[20] = 0$
$q[21] = 0$
$q[22] = 0$
$q[23] = 0$
$q[24] = 0$
$q[25] = 0$
$q[26] = 0$
$q[27] = 0$
$q[28] = 0$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0,02$
$q[32] = 0$

Строка состояния:

$0,55 < R < 0,66$

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)