Тарасов Микита

Лабораторна №1

Мета: 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.

Код для виконання поставленої задачі міститься у файлі **tarasov_lab1.py.** Даний код рахує к-сть заданих літер чи біграм у тексті, їх частоту, і на основі частот – ентропію тексту. Для початку підготовлюємо наш текст (залишаємо лише літери а-я та ‘_’ замість пробілів):

```
5     #prepare text
6     file = "text.txt"
7     with open(file, "r", encoding="utf-8") as file:
8         text = file.read().replace(' ', '_')
9     text = text.lower()
10    text = text.replace('-', '_')
11    text = re.sub(r'\n+', '_', text)
12    text = re.sub(r'_+', '_', text)
13
14    def text_without_spaces(text):
15        return text.replace('_', '')
16    text_without_spaces = text_without_spaces(text)
```

Після цього рахуємо частоту літер, що трапляються в тексті, а далі їх частоти. Так само робимо із біграмами що перетинаються на тексті із пробілами та без, і потім із біграмами, що не перетинаються на тексті із пробілами та без. Результати виконання даного коду наведені у файлі **freqAndEntropy.txt** (нижче наведено лише частину відповідного файлу із результатами)

```
_ : Amount: 112535 [+] Frequency: 0.162586
о : Amount: 64668 [+] Frequency: 0.09343
е : Amount: 48471 [+] Frequency: 0.070029
а : Amount: 47298 [+] Frequency: 0.068334
н : Amount: 38392 [+] Frequency: 0.055467
и : Amount: 38155 [+] Frequency: 0.055125
т : Amount: 35364 [+] Frequency: 0.051092
л : Amount: 30226 [+] Frequency: 0.043669
с : Amount: 30014 [+] Frequency: 0.043363
р : Amount: 26852 [+] Frequency: 0.038795
в : Amount: 23792 [+] Frequency: 0.034374
к : Amount: 20352 [+] Frequency: 0.029404
м : Amount: 18890 [+] Frequency: 0.027291
д : Amount: 18425 [+] Frequency: 0.02662
у : Amount: 16853 [+] Frequency: 0.024348
п : Amount: 16196 [+] Frequency: 0.023399
я : Amount: 11957 [+] Frequency: 0.017275
ы : Amount: 11172 [+] Frequency: 0.016141
ь : Amount: 10642 [+] Frequency: 0.015375
г : Amount: 10237 [+] Frequency: 0.01479
б : Amount: 10081 [+] Frequency: 0.014565
з : Amount: 10011 [+] Frequency: 0.014463
ч : Amount: 8171 [+] Frequency: 0.011805
й : Amount: 6529 [+] Frequency: 0.009433
ж : Amount: 6023 [+] Frequency: 0.008702
х : Amount: 5107 [+] Frequency: 0.007378
ш : Amount: 5077 [+] Frequency: 0.007335
ю : Amount: 3193 [+] Frequency: 0.004613
щ : Amount: 2167 [+] Frequency: 0.003131
э : Amount: 2151 [+] Frequency: 0.003108
ц : Amount: 1906 [+] Frequency: 0.002754
ф : Amount: 1079 [+] Frequency: 0.001559
ъ : Amount: 170 [+] Frequency: 0.000246
ё : Amount: 2 [+] Frequency: 3e-06

Entropy: 4.379028773060594
```

Відповідно, частоти наведені у файлі, згаданому вище. Ентропія також обрахована у цьому файлі, але для подальшого обрахування надлишковості запишемо результати нижче:

	Ентропія	**Надлишковість**
H1 (із пробілом)	4.379028773060594	0.1392504760606287
H1 (без пробілу)	4.464420005196533	0.1149732663024602
H2 (із пробілом, з перетином)	3.9903784772111153	0.2156442552450309
H1 (без пробілів, з перетином)	4.162145373626576	0.1748961968391469
H1 (із пробілом, без перетину)	3.9903267703185494	0.2156544188419076
H1 (без пробілів, без перетину)	4.160968545981837	0.1751294911809283

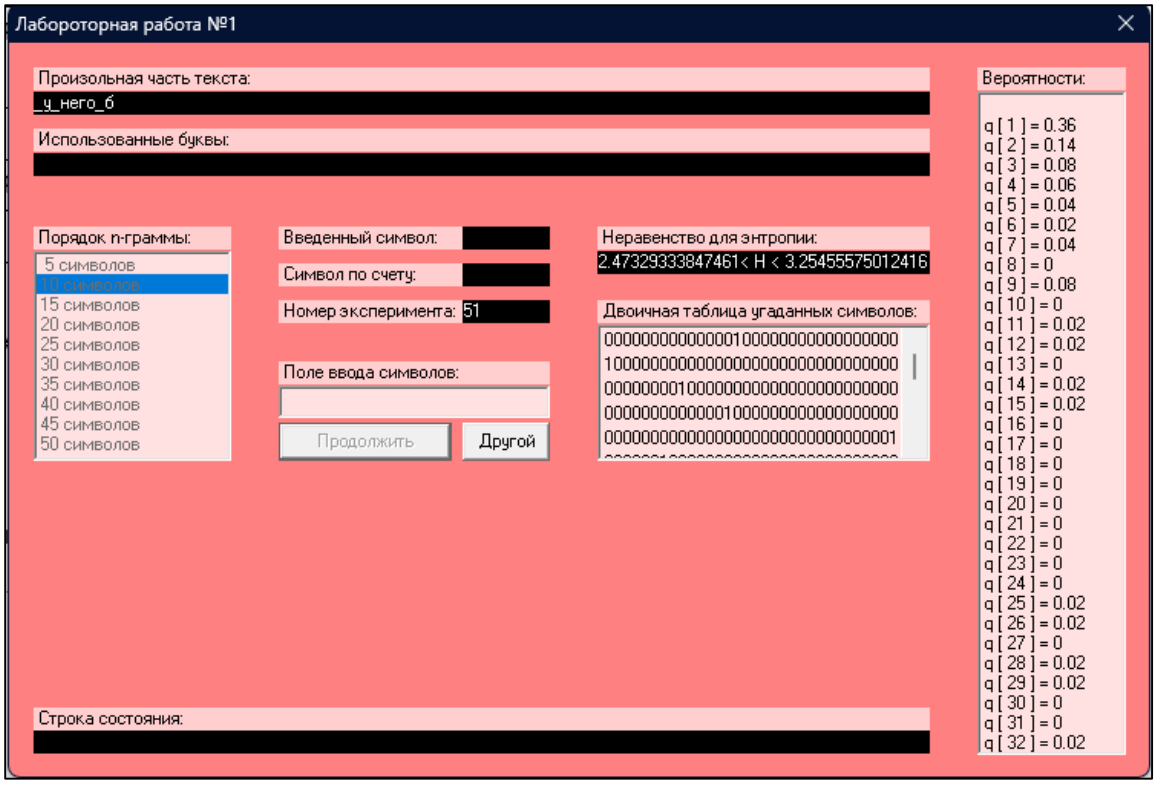
Надлишковість обчислюється за формулою

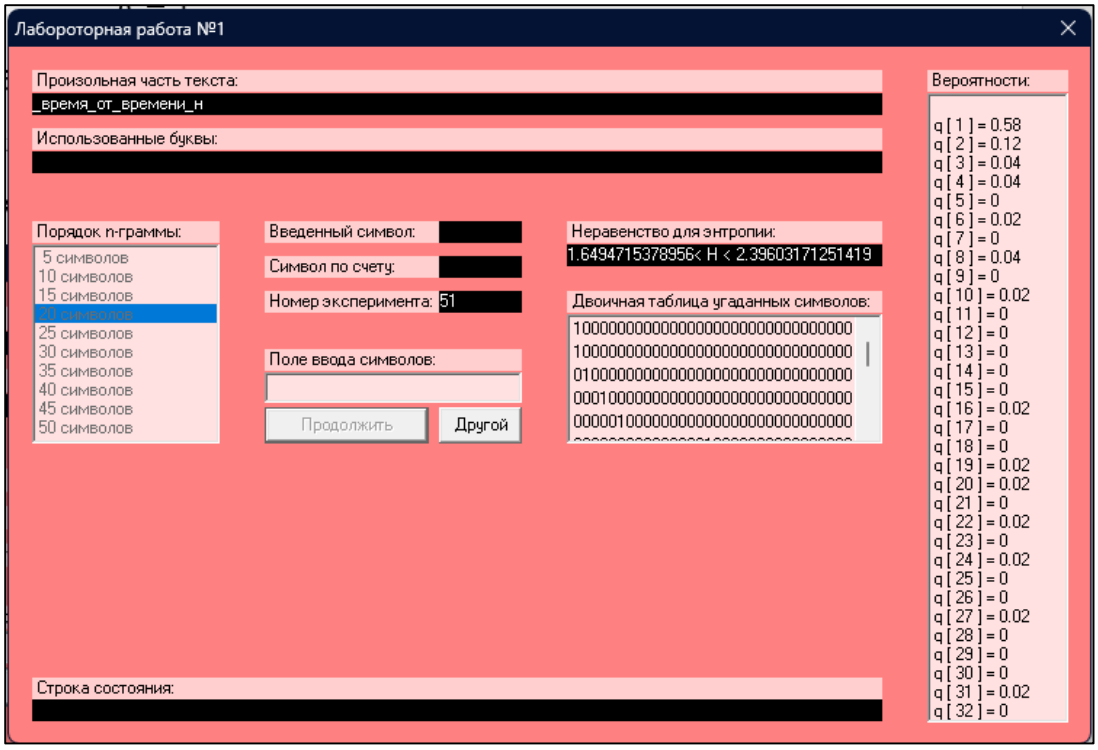
$$R = 1 - \frac{H_\infty}{H_0}$$

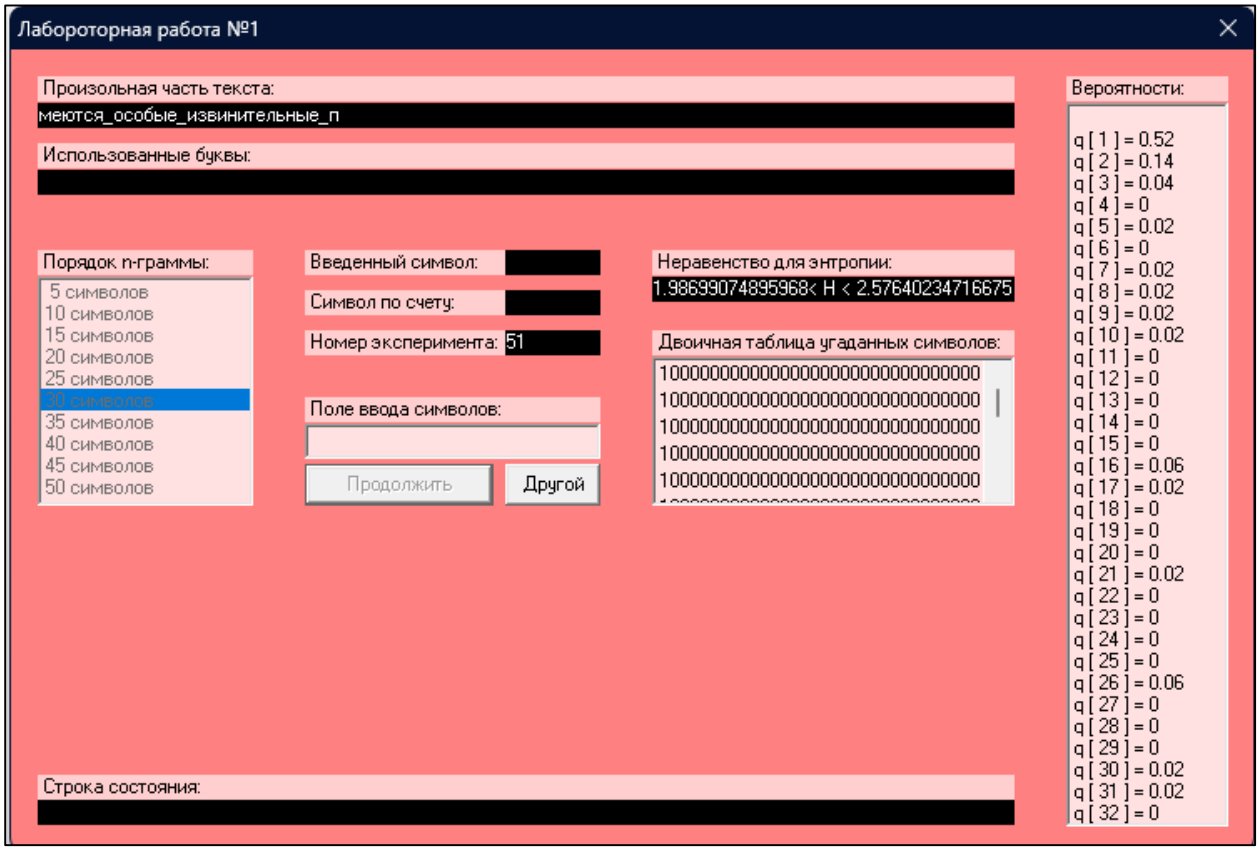
Де H_0 це 5.08746 для тексту із пробілами (log2(34)), 5.04439 для тексті без пробілів (log2(33)).

Переходимо до програми **CoolPinkProgram**

Результати виконання наведені нижче (**H10, H20 та H30 відповідно**):




Лабороторная работа №1
Произольная часть текста:
_время_от_времени_н
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
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1.6494715378956< H < 2.39603171251419
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0.58
q [2] = 0.12
q [3] = 0.04
q [4] = 0.04
q [5] = 0
q [6] = 0.02
q [7] = 0
q [8] = 0.04
q [9] = 0
q [10] = 0.02
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0.02
q [17] = 0
q [18] = 0
q [19] = 0.02
q [20] = 0.02
q [21] = 0
q [22] = 0.02
q [23] = 0
q [24] = 0.02
q [25] = 0
q [26] = 0
q [27] = 0.02
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0.02
q [32] = 0


Лабороторная работа №1
Произольная часть текста:
меются_особые_извинительные_п
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
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1.98699074895968< H < 2.57640234716675
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0.52
q [2] = 0.14
q [3] = 0.04
q [4] = 0
q [5] = 0.02
q [6] = 0
q [7] = 0.02
q [8] = 0.02
q [9] = 0.02
q [10] = 0.02
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0.06
q [17] = 0.02
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0.02
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0.06
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0.02
q [31] = 0.02
q [32] = 0

Отримуємо наступне:

	Ентропія	**Надлишковість**
H(10)	2.47329333847461 < H < 3.25455575012416	0.349088849975168 < R < 0.505341332305078
H(20)	1.6494715378956 < H < 2.39603171251419	0.520793657497162 < R < 0.67010569242088
H(30)	1.98699074895968 < H < 2.57640234716675	0.48471953056665 < R < 0.602601850208064

Висновки:

Під час виконання даного практикуму, я навчився підготовлювати текст для подальшого його аналізу, а таком ознайомвися із поняттям ентропії та навчився обраховувати частоти літер та біграм, а також, користуючись цими значеннями, обраховувати ентропію та надлишковість у різних текстах.