

# КРИПТОГРАФІЯ

## КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

**ФБ-13 Владислав Садохін та Данило Розумовський**

### Мета роботи

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.

### Порядок виконання роботи

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку  $H_1$  та  $H_2$  за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення  $H_1$  та  $H_2$  на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення  $H_1$  та  $H_2$  на тому ж тексті, в якому вилучено всі пробіли.

```
1 import math
2 from collections import Counter
3
4 def is_letter_in_alf(letter):
5     return letter in 'абвгдеёжзийклмнопрстуфхцчшщъыьэюя'
6
7 def is_letter_in_bigram(bigram):
8     return all(is_letter_in_alf(letter) for letter in bigram)
9
10 def calculate_letter_frequencies(text):
11     text = text.lower()
12     letter_count = Counter(letter for letter in text if is_letter_in_alf(letter) or letter == ' ')
13     total_letters_count = sum(letter_count.values())
14     letter_frequencies = {letter: count / total_letters_count for letter, count in letter_count.items()}
15     return letter_frequencies
16
17 def calculate_bigram_frequencies(text, step = 1):
18     text = text.lower()
19     bigram_count = Counter([text[i:i + 2] for i in range(0, len(text) - 1, step) if is_letter_in_bigram(text[i:i + 2])])
20     total_bigram_count = sum(bigram_count.values())
21     bigram_frequencies = {bigram: count / total_bigram_count for bigram, count in bigram_count.items()}
22     return bigram_frequencies
23
24 def calculate_entropy(probabilities):
25     return -sum(p * (0 if p == 0 else math.log2(p)) for p in probabilities.values())
26
```

```

26 |
27 def remove_spaces(text):
28     return ''.join(text.split())
29
30 # Зчитування тексту з файлу
31 file_path = r'C:\Users\alexnd\Desktop\tekstnew.txt'
32
33 with open(file_path, 'r', encoding='utf-8') as file:
34     text = file.read()
35
36 # Видалення пробілів з тексту
37 text_without_spaces = remove_spaces(text)
38
39 # Підрахунок частот букв та біграм та виведення їх
40 letter_probabilities = calculate_letter_frequencies(text)
41 letter_probabilities_no_spaces = calculate_letter_frequencies(text_without_spaces)
42
43 non_overlapping_bigram_probabilities_no_spaces = calculate_bigram_frequencies(text_without_spaces, 2)
44 non_overlapping_bigram_probabilities = calculate_bigram_frequencies(text, 2)
45
46 overlapping_bigram_probabilities_no_spaces = calculate_bigram_frequencies(text_without_spaces)
47 overlapping_bigram_probabilities = calculate_bigram_frequencies(text)
48
49 print("Частоти букв з пробілами:")
50 for letter, probability in letter_probabilities.items():

```

```

50 for letter, probability in letter_probabilities.items():
51     print(f"{letter}: {probability:.6f}")
52
53 print("Частоти букв без пробілів:")
54 for letter, probability in letter_probabilities_no_spaces.items():
55     print(f"{letter}: {probability:.6f}")
56
57 print("\nЧастоти біграм з перетином та пробілами:")
58 for bigram, probability in overlapping_bigram_probabilities.items():
59     print(f"{bigram}: {probability:.6f}")
60
61 print("Частоти біграм з перетином без пробілів:")
62 for bigram, probability in overlapping_bigram_probabilities_no_spaces.items():
63     print(f"{bigram}: {probability:.6f}")
64
65 print("\nЧастоти біграм без перетину з пробілами:")
66 for bigram, probability in non_overlapping_bigram_probabilities.items():
67     print(f"{bigram}: {probability:.6f}")
68
69 print("Частоти біграм без перетину без пробілів:")
70 for bigram, probability in non_overlapping_bigram_probabilities_no_spaces.items():
71     print(f"{bigram}: {probability:.6f}")
72
73 # Підрахунок H1 і H2 за безпосереднім означенням
74 H1 = calculate_entropy(letter_probabilities)

```

```

73 # Підрахунок H1 і H2 за безпосереднім означенням
74 H1 = calculate_entropy(letter_probabilities)
75 H1_no_spaces = calculate_entropy(letter_probabilities_no_spaces)
76
77 H2_overlapping_no_spaces = calculate_entropy(overlapping_bigram_probabilities_no_spaces)/2
78 H2_overlapping = calculate_entropy(overlapping_bigram_probabilities)/2
79 H2_non_overlapping = calculate_entropy(non_overlapping_bigram_probabilities)/2
80 H2_non_overlapping_no_spaces = calculate_entropy(non_overlapping_bigram_probabilities_no_spaces)/2
81
82 print(f"\nH1 (Ентропія букв з пробілами): {H1}")
83 print(f"H1 (Ентропія букв без пробілів): {H1_no_spaces}")
84 print(f"H2 (Ентропія біграм з перетином та пробілами): {H2_overlapping}")
85 print(f"H2 (Ентропія біграм з перетином без пробілів): {H2_overlapping_no_spaces}")
86 print(f"H2 (Ентропія біграм без перетину з пробілами): {H2_non_overlapping}")
87 print(f"H2 (Ентропія біграм без перетину без пробілів): {H2_non_overlapping_no_spaces}")

```

Текст був взятий з книги ,Робінзон Крузо, Даниель Дефо перекладеної на рос.мову.

|                           |
|---------------------------|
| Частоти букв з пробілами: |
| е: 0.071370               |
| с: 0.043650               |
| л: 0.041818               |
| и: 0.056086               |
| : 0.169165                |
| у: 0.022645               |
| щ: 0.002502               |
| т: 0.049919               |
| в: 0.034276               |
| н: 0.051917               |
| а: 0.059080               |
| о: 0.097866               |
| р: 0.034909               |
| я: 0.023572               |
| п: 0.023700               |
| к: 0.027598               |
| ю: 0.004436               |
| ч: 0.012491               |
| й: 0.008666               |
| г: 0.015183               |
| ц: 0.002264               |
| з: 0.013273               |
| ж: 0.008418               |
| ь: 0.017921               |
| б: 0.017453               |
| м: 0.029682               |
| д: 0.027138               |
| ы: 0.015604               |
| э: 0.002321               |
| х: 0.007937               |
| ш: 0.006555               |
| ф: 0.000345               |
| ъ: 0.000238               |

|                                   |
|-----------------------------------|
| Частоти букв <b>без</b> пробілів: |
| е: 0.085902                       |
| с: 0.052538                       |
| л: 0.050332                       |
| и: 0.067506                       |
| у: 0.027256                       |
| щ: 0.003012                       |
| т: 0.060083                       |
| в: 0.041255                       |
| н: 0.062488                       |
| а: 0.071109                       |
| о: 0.117792                       |
| р: 0.042017                       |
| я: 0.028371                       |
| п: 0.028525                       |
| к: 0.033217                       |
| ю: 0.005339                       |
| ч: 0.015035                       |
| й: 0.010431                       |
| г: 0.018274                       |
| ц: 0.002725                       |
| з: 0.015975                       |
| ж: 0.010132                       |
| ь: 0.021570                       |
| б: 0.021007                       |
| м: 0.035725                       |
| д: 0.032664                       |
| ы: 0.018782                       |
| э: 0.002794                       |
| х: 0.009553                       |
| ш: 0.007889                       |
| ф: 0.000415                       |
| ъ: 0.000286                       |

|   |   |
|---|---|
| Частоти біграм з перетином <b>та</b> пробілами: | Частоти біграм з перетином <b>без</b> пробілів: |
| ес: 0.007359                                    | ес: 0.007794                                    |
| сл: 0.004622                                    | сл: 0.003726                                    |
| ли: 0.010647                                    | ли: 0.009107                                    |
| су: 0.001380                                    | ис: 0.004986                                    |
| ущ: 0.000371                                    | су: 0.001147                                    |
| ще: 0.001973                                    | ущ: 0.000295                                    |
| ст: 0.017542                                    | ще: 0.001571                                    |
| тв: 0.003167                                    | ст: 0.014250                                    |
| ву: 0.000894                                    | тв: 0.002897                                    |
| уе: 0.000215                                    | ву: 0.000799                                    |
| ет: 0.004498                                    | уе: 0.000239                                    |
| на: 0.013901                                    | ет: 0.004717                                    |
| св: 0.002450                                    | тн: 0.002047                                    |
| ве: 0.007581                                    | на: 0.011094                                    |
| те: 0.006076                                    | ас: 0.005801                                    |

|   |   |
|---|---|
| Частоти біграм <b>без</b> перетину з пробілами: | Частоти біграм <b>без</b> перетину <b>без</b> пробілів: |
| ес: 0.007456                                    | ес: 0.007957  |
| ли: 0.010523                                    | ли: 0.009091  |
| ущ: 0.000399                                    | су: 0.001170  |
| тв: 0.003218                                    | ще: 0.001598  |
| уе: 0.000204                                    | ст: 0.014150  |
| на: 0.013896                                    | ву: 0.000809  |
| ве: 0.007518                                    | ет: 0.004797  |
| те: 0.006067                                    | на: 0.011004  |
| ст: 0.017442                                    | св: 0.002050  |
| ор: 0.008814                                    | еи: 0.001534  |
| ия: 0.002144                                    | ор: 0.007402  |
| ри: 0.006325                                    | ия: 0.002860  |



Произвольная часть текста:  
ятя\_между\_собой\_ин

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:  
 $2,19010643975419 < H < 2,797780939298$

Двоичная таблица угаданных символов:

|                                  |
|----------------------------------|
| 10000000000000000000000000000000 |
| 00000000000000000000000000000000 |
| 10000000000000000000000000000000 |
| 00000000000000000000000000000000 |
| 00000000000000000000000000000000 |
| 00000000000000000000000000000000 |
| 10000000000000000000000000000000 |

Вероятности:

|              |
|--------------|
| q[1] = 0,48  |
| q[2] = 0,12  |
| q[3] = 0,06  |
| q[4] = 0,04  |
| q[5] = 0     |
| q[6] = 0     |
| q[7] = 0     |
| q[8] = 0,04  |
| q[9] = 0     |
| q[10] = 0    |
| q[11] = 0,02 |
| q[12] = 0    |
| q[13] = 0    |
| q[14] = 0    |
| q[15] = 0,02 |
| q[16] = 0    |
| q[17] = 0,04 |
| q[18] = 0,04 |
| q[19] = 0,04 |
| q[20] = 0    |
| q[21] = 0    |
| q[22] = 0    |
| q[23] = 0,02 |
| q[24] = 0    |
| q[25] = 0,02 |
| q[26] = 0,04 |
| q[27] = 0    |
| q[28] = 0    |
| q[29] = 0,02 |
| q[30] = 0    |
| q[31] = 0    |
| q[32] = 0    |

Строка состояния:

Результаты для  $H^{(20)}$ 

Произвольная часть текста:  
бходимости\_выполнить\_обещание\_фактически\_выглядит\_так\_что\_обе\_стороны\_имели

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: \_ (пробел)

Символ по счету: 1

Номер эксперимента: 51

Неравенство для энтропии:  
 $1,58457439647096 < H < 2,321808697683$

Двоичная таблица угаданных символов:

|                                  |
|----------------------------------|
| 01000000000000000000000000000000 |
| 00000000000000000000000000000000 |
| 10000000000000000000000000000000 |
| 10000000000000000000000000000000 |
| 10000000000000000000000000000000 |
| 00000000000000000000000000000000 |

Вероятности:

|                           |
|---------------------------|
| q[1] = 0,568627450980392  |
| q[2] = 0,156862745098039  |
| q[3] = 0,0392156862745098 |
| q[4] = 0,0196078431372549 |
| q[5] = 0,0196078431372549 |
| q[6] = 0,0196078431372549 |
| q[7] = 0                  |
| q[8] = 0                  |
| q[9] = 0,0196078431372549 |
| q[10] = 0                 |
| q[11] = 0,019607843137254 |
| q[12] = 0                 |
| q[13] = 0                 |
| q[14] = 0,019607843137254 |
| q[15] = 0,039215686274509 |
| q[16] = 0                 |
| q[17] = 0                 |
| q[18] = 0                 |
| q[19] = 0,019607843137254 |
| q[20] = 0                 |
| q[21] = 0                 |
| q[22] = 0,039215686274509 |
| q[23] = 0                 |
| q[24] = 0,019607843137254 |
| q[25] = 0                 |
| q[26] = 0                 |
| q[27] = 0                 |
| q[28] = 0                 |
| q[29] = 0                 |
| q[30] = 0                 |
| q[31] = 0                 |
| q[32] = 0                 |

Строка состояния:  
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

Результаты для  $H^{(30)}$

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

|   |
|---|
| N1 (Ентропія букв з пробілами): 4.360373178797449                 |
| N1 (Ентропія букв без пробілів): 4.458867100342456                |
| N2 (Ентропія біграм з перетином та пробілами): 3.9195543933732493 |
| N2 (Ентропія біграм з перетином без пробілів): 4.118281456606397  |
| N2 (Ентропія біграм без перетину з пробілами): 3.919944391193998  |
| N2 (Ентропія біграм без перетину без пробілів): 4.116656741575615 |

Надлишковість джерела відкритого тексту (мови) дорівнює  $R = 1 - \frac{H_{\infty}}{H_0}$

Надлишковість для нашого файлу

|                        | з пробілами | без пробілів |
|------------------------|-------------|--------------|
| R <sub>1</sub>         | 0,128       | 0,108        |
| R <sub>2</sub> (крок1) | 0,2161      | 0,176        |
| R <sub>2</sub> (крок2) | 0,2160      | 0,177        |

Де крок1- з перетином, крок2 – без перетину

Надлишковість згідно з експериментами у CoolPinkProgram.exe

| CoolPinkProgram   |       |       |
|-------------------|-------|-------|
|                   | min   | max   |
| H <sub>(10)</sub> | 2,431 | 2,983 |
| H <sub>(20)</sub> | 2,190 | 2,797 |
| H <sub>(30)</sub> | 1,587 | 2,321 |

| CoolPinkProgram   |       |       |
|-------------------|-------|-------|
|                   | min   | max   |
| R <sub>(10)</sub> | 0,514 | 0,403 |
| R <sub>(20)</sub> | 0,562 | 0,441 |
| R <sub>(30)</sub> | 0,683 | 0,536 |

## Висновки

У ході виконання лабораторної роботи було обчислено частоти букв, біграм та ентропії, для випадків тексту відкритого джерела з пробілами та без, також з перетинами та без. Засвоєно ключові поняття ентропії на символ джерела та його надлишковості, пораховано надлишковість рос.мови у тексті, який ми знайшли самі та за допомогою програми CoolPinkProgram.exe, в якій містився інший текст.