# Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського" Фізико-технічний інститут

# Криптографія

Комп'ютерний практикум №1 Експериментальна оцінка на символ джерела відкритого тексту

Виконали:

Студенти 3 курсу

Загородній Я.М, Венгер П.Ю.

Перевірив:

**Мета роботи:** засвоєння поняття 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

#### Постановка задачі:

- 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.
- 2. За допомогою програми CoolPinkProgram оцінити значення  $\mathbf{H}^{(10)}, \mathbf{H}^{(20)}, \mathbf{H}^{(30)}$
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

### Хід роботи

Пункт 1. Весь код програми поданий в lab1\_cryptography.ipynb, також будуть прикріплені файли excel з табличками

```
frequency= sorted(freq, key=lambda x : alpcount[x],reverse=1)
  relative_frequencies = []
 v for letter in frequency:
      relative_frequencies.append(freq[letter])
   frequency_data = pd.DataFrame(index=frequency)
   frequency_data['relative_frequency'] = relative_frequencies
  # Виводимо перші кілька рядків DataFrame print(frequency_data.head(10))
   frequency_data.to_excel("relative_frequencies.xlsx")
✓ 0.1s
  relative_frequency
           0.110161
            0.081928
           0.079719
            0.064138
            0.063714
            0.056262
            0.051702
            0.048350
            0.044482
            0.043574
```

```
from collections import Counter
  input_text = string_text
  for i in range(len(input_text) - 1):
    bigrams.append(input_text[i] + input_text[i + 1])
  double_bigrams = []
  for i in range(0, len(input_text) - 2, 2):
      double_bigrams.append(input_text[i] + input_text[i + 1])
   bigram_counts = dict(Counter(bigrams))
   bigram_frequencies = {bigram: count / len(bigrams) for bigram, count in bigram_counts.items()}
  double_bigram_counts = dict(Counter(double_bigrams))
  double_bigram_frequencies = {bigram: count / len(double_bigrams) for bigram, count in double_bigram_counts.items()}
   entropy_contributions = []
   for frequency in bigram_frequencies.values():
      if frequency > 0: -- # Переконайтеся, що частота не дорівнює нулю, щоб уникнути помилок
         entropy_contributions.append(frequency * math.log(frequency, 2))
   print(entropy_contributions)
  H1 = -sum(entropy_contributions)
   print("Entropy H1:", H1)
Entropy H1: 8.119527189870336
```

```
# dynada pan commune emponis

# def calculate_entropy(blgnum, freq blgnum);

# entropy_contributions / 1

# of frequency in freq blgnum values();

# entropy_contributions / 2

# dynamine emponis an element is an emponis emponis emponis emponis emponis pan element is an emponis pan element is allement is an emponis pan element is an emponis pan element is allement is an emponis pan element is allement is an emponis pan element is an emponis pan element is allement is an emponis pan element is
```

| En  | Entropy of bigrams (H2): 4.238694263913949         |          |          |          |          |          |          |   |
|-----|--|----------|----------|----------|----------|----------|----------|---|
| En  | Entropy of double bigrams (H22): 4.238452024542541 |          |          |          |          |          |          |   |
|     | a  | 6        | В        | г        | д        | e        | ë        |   |
| а   | 0.000327   | 0.001507 | 0.005468 | 0.001566 | 0.003021 | 0.00172  | 0        | \ |
| 6   | 0.001235   | 0.000024 | 0.0001   | 0.00001  | 0.000026 | 0.002384 | 0        |   |
| В   | 0.006547   | 0.00025  | 0.000466 | 0.000428 | 0.000746 | 0.005195 | 0        |   |
| г   | 0.001125   | 0.000037 | 0.000125 | 0.000011 | 0.001146 | 0.000634 | 0        |   |
| д   | 0.004848   | 0.000085 | 0.001061 | 0.000062 | 0.000074 | 0.005134 | 0.000001 |   |
| e   | 0.000212   | 0.002265 | 0.003504 | 0.004431 | 0.0038   | 0.002177 | 0        |   |
| ë   | 0.000002   | 0.000003 | 0.000006 | 0.000002 | 0.000002 | 0.000002 | 0        |   |
| ж   | 0.001429   | 0.000078 | 0.000027 | 0.000025 | 0.000808 | 0.00414  | 0        |   |
| 3   | 0.006049   | 0.000201 | 0.000952 | 0.000534 | 0.000931 | 0.000293 | 0        |   |
| И   | 0.000344   | 0.001356 | 0.004964 | 0.001221 | 0.002802 | 0.003322 | 0        |   |
| й   | 0.000202   | 0.000396 | 0.000768 | 0.000354 | 0.000761 | 0.000122 | 0        |   |
| κ   | 0.008325   | 0.000503 | 0.000608 | 0.000171 | 0.000253 | 0.000617 | 0.000002 |   |
| л   | 0.008195   | 0.00033  | 0.000798 | 0.000397 | 0.000605 | 0.004843 | 0.000004 |   |
| м   | 0.003418   | 0.000328 | 0.00079  | 0.000365 | 0.00039  | 0.003354 | 0        |   |
| н   | 0.012474   | 0.000342 | 0.000515 | 0.000301 | 0.001386 | 0.010396 | 0        |   |
| o   | 0.000211   | 0.00504  | 0.011267 | 0.005852 | 0.006169 | 0.003234 | 0.000001 |   |
| п   | 0.001226   | 0.000001 | 0.000001 | 0.000015 | 0.000002 | 0.002834 | 0        |   |
| р   | 0.009495   | 0.000253 | 0.000496 | 0.000397 | 0.00049  | 0.006863 | 0        |   |
| c   | 0.001734   | 0.000201 | 0.002301 | 0.000162 | 0.000532 | 0.004041 | 0.000038 |   |
| т   | 0.006594   | 0.000274 | 0.003256 | 0.000139 | 0.000399 | 0.005324 | 0.000004 |   |
| у   | 0.000173   | 0.000863 | 0.001574 | 0.001578 | 0.002297 | 0.000341 | 0        |   |
| ф   | 0.000195   | 0.000009 | 0.000028 | 0.000006 | 0.000005 | 0.000188 | 0        |   |
| ••• |  |          |          |          |          |          |          |   |
| ю   | 0  | 0        | 0        | 0.000049 | 0.000045 | 0.000044 |          |   |
| я   | 0  | 0        | 0        | 0.000175 | 0.000128 | 0.000285 |          |   |
|     |  |          |          |          |          |          |          |   |

```
# Створення DataFrame з алфавіту
   frequency_df = pd.DataFrame(index=alphabet, columns=alphabet)
   # Генерація всіх можливих біграм з алфавіту
   bigrams_list = [letter1 + letter2 for letter1 in alphabet for letter2 in alphabet]
   # Заповнення DataFrame біграмами
 v for idx, letter in enumerate(alphabet):
      frequency_df[letter] = bigrams_list[idx * len(alphabet):(idx + 1) * len(alphabet)]
   # Транспонування DataFrame для правильного формату
   frequency_df = frequency_df.T
   # Заповнення DataFrame частотами двійних біграм
 v for bigram in double_bigram_frequencies.keys():
       row_idx, col_idx = np.where(frequency_df == bigram)
       frequency_df.iloc[row_idx, col_idx] = double_bigram_frequencies[bigram]
   # Встановлення 0 для біграм, які не мають частоти

√ for bigram in bigrams list:

       row_idx, col_idx = np.where(frequency_df == bigram)
      frequency_df.iloc[row_idx, col_idx] = 0
   # Вивід фінального DataFrame
   print(frequency_df.head())
   # Збереження DataFrame у файл Excel
   frequency df.to excel("double bigram frequencies.xlsx")

√ 0.7s

                 6
                                               Д
                                                        еë
                           В
                                                                     ж
 0.00032 0.001537 0.005553 0.001611 0.002976 0.001749 0 0.001443 \
6 0.001226 0.00002 0.000125 0.000008 0.000028 0.002376 0 0.000008
в 0.006535 0.000274 0.000464 0.000428 0.000756 0.005116 0 0.000053
г 0.001096 0.000036 0.000116 0.000011 0.001104 0.000632 0 0.000005
д 0.004786 0.000083 0.001059 0.000075 0.000066 0.005152 0 0.000024
                            ц
                                                  ш
                                                             щ
a 0.004863 0.001559 ... 0.000136 0.001331 0.001676 0.000265
                                                                       0 \
6 0.000009 0.000889 ... 0.000006 0.00002 0.000008 0.000252
                                                                 0.00011
в 0.000665 0.003762 ... 0.000058 0.000238 0.001247 0.000009 0.000028
 0.000045 0.000941 ... 0.000003 0.000035 0.000009
                                                            a
                                                                       a
д 0.000052 0.002937 ... 0.000202 0.000075 0.000127 0.000002
                                                                 0.00021
                                               я
         ы
                  ь
                            э
                                     ю
                  0 0.000345 0.000949 0.003096
```

Результат

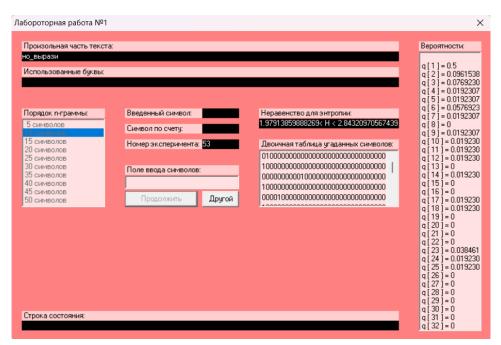
| Текст без пробілів |                                | Текст з пробілами                    |  |
|--------------------|--------------------------------|--------------------------------------|--|
| Букви              | liters.xlsx                    | liters_space.xlsx                    |  |
| Біграма            | relative_frequencies.xlsx      | relative_frequencies_space.xlsx      |  |
| Біграма з кроком 2 | double_bigram_frequencies.xlsx | double_bigram_frequencies_space.xlsx |  |

#### Ентропія

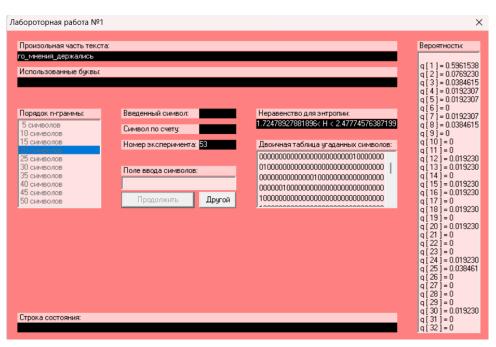
|                | Текст без пробілів | Текст з пробілами  |
|----------------|--------------------|--------------------|
| H1             | 8.477388527827898  | 8.119527189870336  |
| H2             | 4.238694263913949  | 4.059763594935168  |
| Н2(з кроком 2) | 4.238452024542541  | 4.0596450054514275 |

Пункт 2. За допомогою програми CoolPinkProgram оцінити значення  $H^{(10)}$ ,  $H^{(20)}$ ,  $H^{(30)}$ .

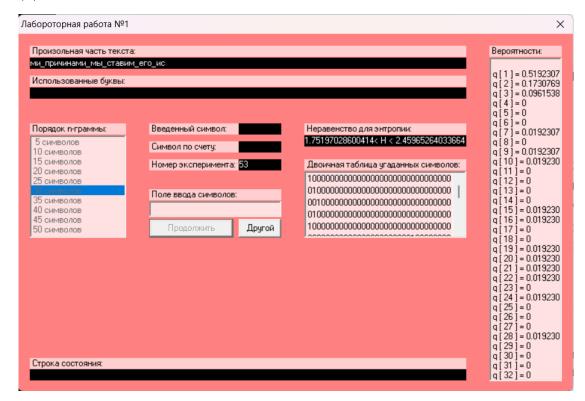
## Для H<sup>(10)</sup>:



# Для H<sup>(20)</sup>:



# Для H<sup>(30)</sup>:



|            | Ентропія              | Надлишковість         |
|------------|-----------------------|-----------------------|
| $H^{(10)}$ | 1.97914 < H < 2.84321 | 0.43136 < H < 0.60417 |
| $H^{(20)}$ | 1.72479 < H < 2.47775 | 0.50445 < H < 0.65504 |
| $H^{(30)}$ | 1.75197 < H < 2.45965 | 0.50807 < R < 0.64961 |

#### Висновок

Ознайомилися з поняттями ентропії на символ джерела та його надлишковості, навчилися визначати частоти літер та біграм на довільному тексті.

3 отриманих значень можемо визначити, що ентропія H1 з пробілами більша за H1 без пробілів. Ентропія та надлишковість перехресних та неперехресних біграм без пробілів (або з пробілами) (H2 та H2(з кроком 2)) майже не відрізняються.