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

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

з дисципліни «Криптографія»

Тема: «Експериментальна оцінка ентропії на символ джерела відкритого тексту»

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.

Хід роботи

Завдання 1

Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

Виконання

```
import re
import pandas as pd
from collections import Counter
from math import log2

def preprocess_text(text, remove_spaces=False):
    """Очистка тексту: перетворення в нижній регістр, видалення неалфавітних символів."""
    cleaned_text = text.lower()
    cleaned_text = re.sub(r'[^а-яё ]', '', cleaned_text)
    cleaned_text = re.sub(r'\s+', ' ', cleaned_text)
    if remove_spaces:
        cleaned_text = cleaned_text.replace(" ", "")
    return cleaned_text

def calculate_frequencies(text):
    """Обчислення частот літер і біграм."""
    letter_counts = Counter(text)
    total_letters = sum(letter_counts.values())
    letter_frequencies = {char: count / total_letters for char, count in letter_counts.items()}

    bigram_counts_with_overlap = Counter(
        [text[i:i + 2] for i in range(len(text) - 1) if " " not in text[i:i + 2]]
    )
    total_bigrams_with_overlap = sum(bigram_counts_with_overlap.values())
    bigram_frequencies_with_overlap = {
        bigram: count / total_bigrams_with_overlap for bigram, count in bigram_counts_with_overlap.items()
    }

    bigram_counts_no_overlap = Counter(
        [text[i:i + 2] for i in range(0, len(text) - 1, 2) if " " not in text[i:i + 2]]
    )
    total_bigrams_no_overlap = sum(bigram_counts_no_overlap.values())
    bigram_frequencies_no_overlap = {
        bigram: count / total_bigrams_no_overlap for bigram, count in bigram_counts_no_overlap.items()
    }

    return letter_frequencies, bigram_frequencies_with_overlap, bigram_frequencies_no_overlap

def calculate_entropy(frequencies):
```

```
    """Обчислення ентропії."""
    return -sum(freq * log2(freq) for freq in frequencies.values())

def save_frequencies_to_csv(data, filename, index_label=None,
columns_label=None):
    """Збереження частот у CSV-файл."""
    dataframe = pd.DataFrame.from_dict(data, orient='index',
columns=[columns_label])
    dataframe = dataframe.sort_values(by=columns_label, ascending=False)
    dataframe.to_csv(filename, index_label=index_label)

def main():
    with open("input_text.txt", "r", encoding="utf-8") as file:
        raw_text = file.read()

    text_with_spaces = preprocess_text(raw_text, remove_spaces=False)
    text_without_spaces = preprocess_text(raw_text, remove_spaces=True)

    letter_freq_with_spaces, bigram_freq_with_overlap_with_spaces,
bigram_freq_no_overlap_with_spaces = calculate_frequencies(text_with_spaces)
    letter_freq_without_spaces, bigram_freq_with_overlap_without_spaces,
bigram_freq_no_overlap_without_spaces =
calculate_frequencies(text_without_spaces)

    entropy_letters_with_spaces = calculate_entropy(letter_freq_with_spaces)
    entropy_bigrams_with_overlap_with_spaces =
calculate_entropy(bigram_freq_with_overlap_with_spaces) / 2
    entropy_bigrams_no_overlap_with_spaces =
calculate_entropy(bigram_freq_no_overlap_with_spaces) / 2

    entropy_letters_without_spaces =
calculate_entropy(letter_freq_without_spaces)
    entropy_bigrams_with_overlap_without_spaces =
calculate_entropy(bigram_freq_with_overlap_without_spaces) / 2
    entropy_bigrams_no_overlap_without_spaces =
calculate_entropy(bigram_freq_no_overlap_without_spaces) / 2

    print("\nРезультати для тексту з пробілами:")
    print(f"Ентропія літер: {entropy_letters_with_spaces:.4f}")
    print(f"Ентропія біграм (з перетином):
{entropy_bigrams_with_overlap_with_spaces:.4f}")
    print(f"Ентропія біграм (без перетину):
{entropy_bigrams_no_overlap_with_spaces:.4f}")

    print("\nРезультати для тексту без пробілів:")
    print(f"Ентропія літер: {entropy_letters_without_spaces:.4f}")
    print(f"Ентропія біграм (з перетином):
{entropy_bigrams_with_overlap_without_spaces:.4f}")
    print(f"Ентропія біграм (без перетину):
{entropy_bigrams_no_overlap_without_spaces:.4f}")

    save_frequencies_to_csv(letter_freq_with_spaces,
"letter_frequencies_with_spaces.csv", index_label="Літера",
columns_label="Частота")
    save_frequencies_to_csv(bigram_freq_with_overlap_with_spaces,
"bigram_frequencies_with_overlap_with_spaces.csv", index_label="Біграма",
columns_label="Частота з перетином")
    save_frequencies_to_csv(bigram_freq_no_overlap_with_spaces,
"bigram_frequencies_no_overlap_with_spaces.csv", index_label="Біграма",
columns_label="Частота без перетину")

    save_frequencies_to_csv(letter_freq_without_spaces,
"letter_frequencies_without_spaces.csv", index_label="Літера",
columns_label="Частота")
    save_frequencies_to_csv(bigram_freq_with_overlap_without_spaces,
"bigram_frequencies_with_overlap_without_spaces.csv", index_label="Біграма",
```

```
columns_label="Частота з перетином")
    save_frequencies_to_csv(bigram_freq_no_overlap_without_spaces,
"bigram_frequencies_no_overlap_without_spaces.csv", index_label="Біграма",
columns_label="Частота без перетину")

if __name__ == "__main__":
    main()
```

```
Результати для тексту з пробілами:
Ентропія літер: 4.3586
Ентропія біграм (з перетином): 3.9451
Ентропія біграм (без перетину): 3.9450
```

```
Результати для тексту без пробілів:
Ентропія літер: 4.4515
Ентропія біграм (з перетином): 4.1280
Ентропія біграм (без перетину): 4.1265
```

```
Файл: bigram_frequencies_no_overlap_without_spaces.csv
  Біграма  Частота без перетину
0      то              0.018094
1      ов              0.012867
2      на              0.012257
3      не              0.012074
4      но              0.011844
```

```
Файл: bigram_frequencies_no_overlap_with_spaces.csv
  Біграма  Частота без перетину
0      то              0.022192
1      не              0.015329
2      на              0.015023
3      но              0.014430
4      ст              0.014263
```

```
Файл: bigram_frequencies_with_overlap_without_spaces.csv
  Біграма  Частота з перетином
0      то             0.018081
1      ов             0.012588
2      не             0.012235
3      на             0.012100
4      но             0.011900
```

```
Файл: bigram_frequencies_with_overlap_with_spaces.csv
  Біграма  Частота з перетином
0      то             0.022072
1      не             0.015184
2      на             0.015065
3      но             0.014491
4      ст             0.014133
```

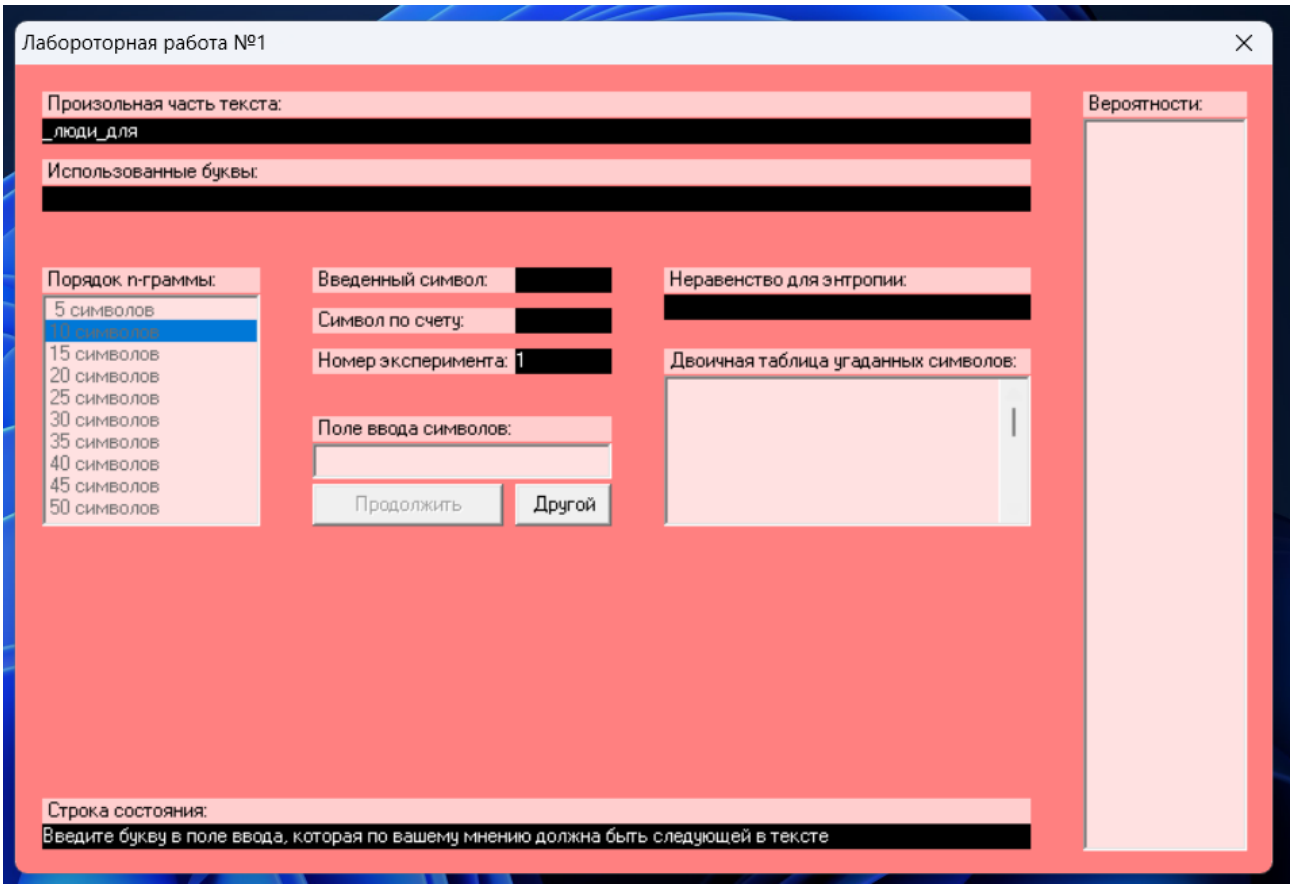
```
Файл: letter_frequencies_without_spaces.csv
  Літера   Частота
0      о  0.114725
1      е  0.087090
2      а  0.079660
3      н  0.065086
4      и  0.064848

Файл: letter_frequencies_with_spaces.csv
  Літера   Частота
0         0.167166
1      о  0.095547
2      е  0.072531
3      а  0.066344
4      н  0.054206
```

Завдання 2

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

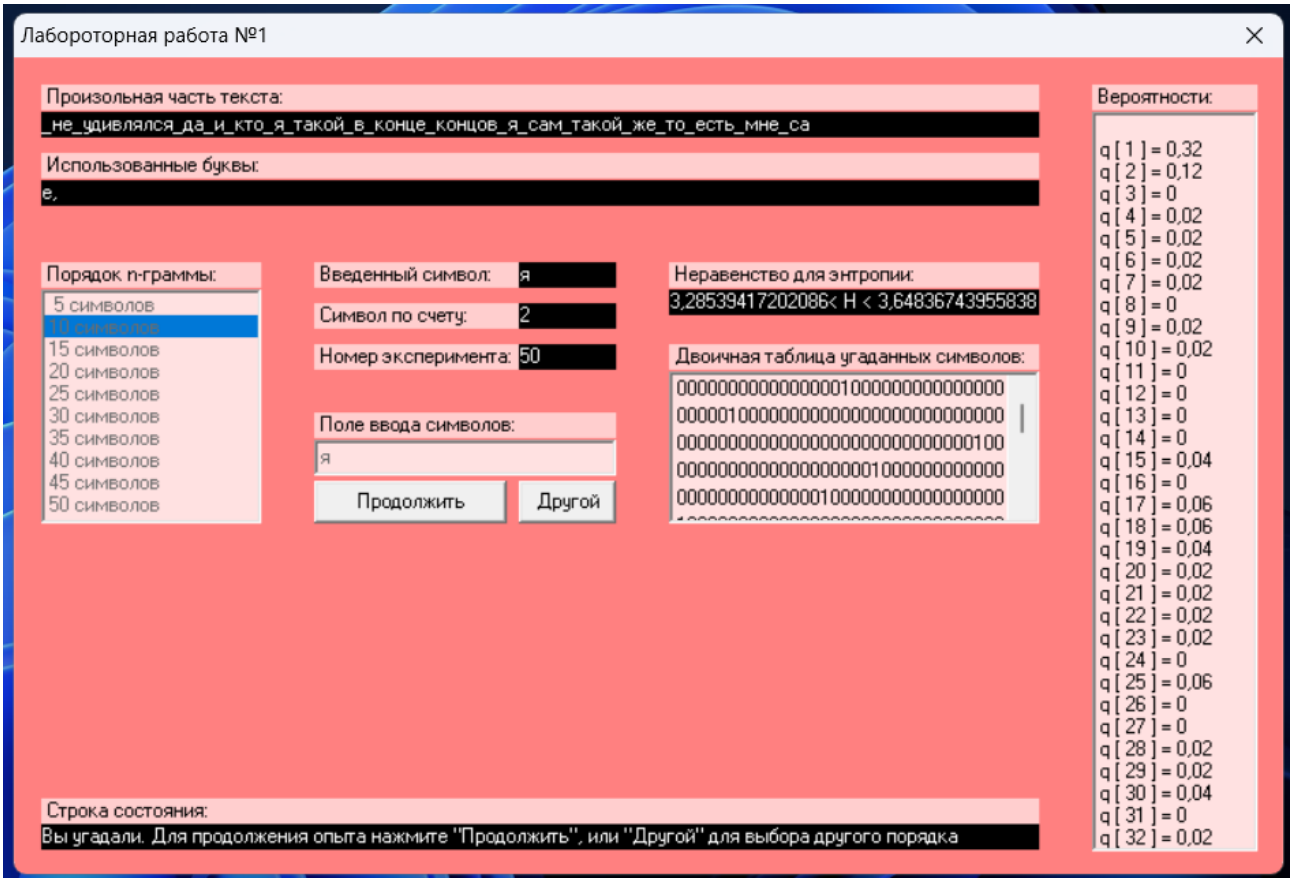
Під час виконання були проблеми з ієроглєфами, і на жаль, повністю прибрати її не вдалось, але залишок символів не заважав виконання роботи, оскільки одна кнопка відповідала за продовження дії, а інша до перехід до наступного значення

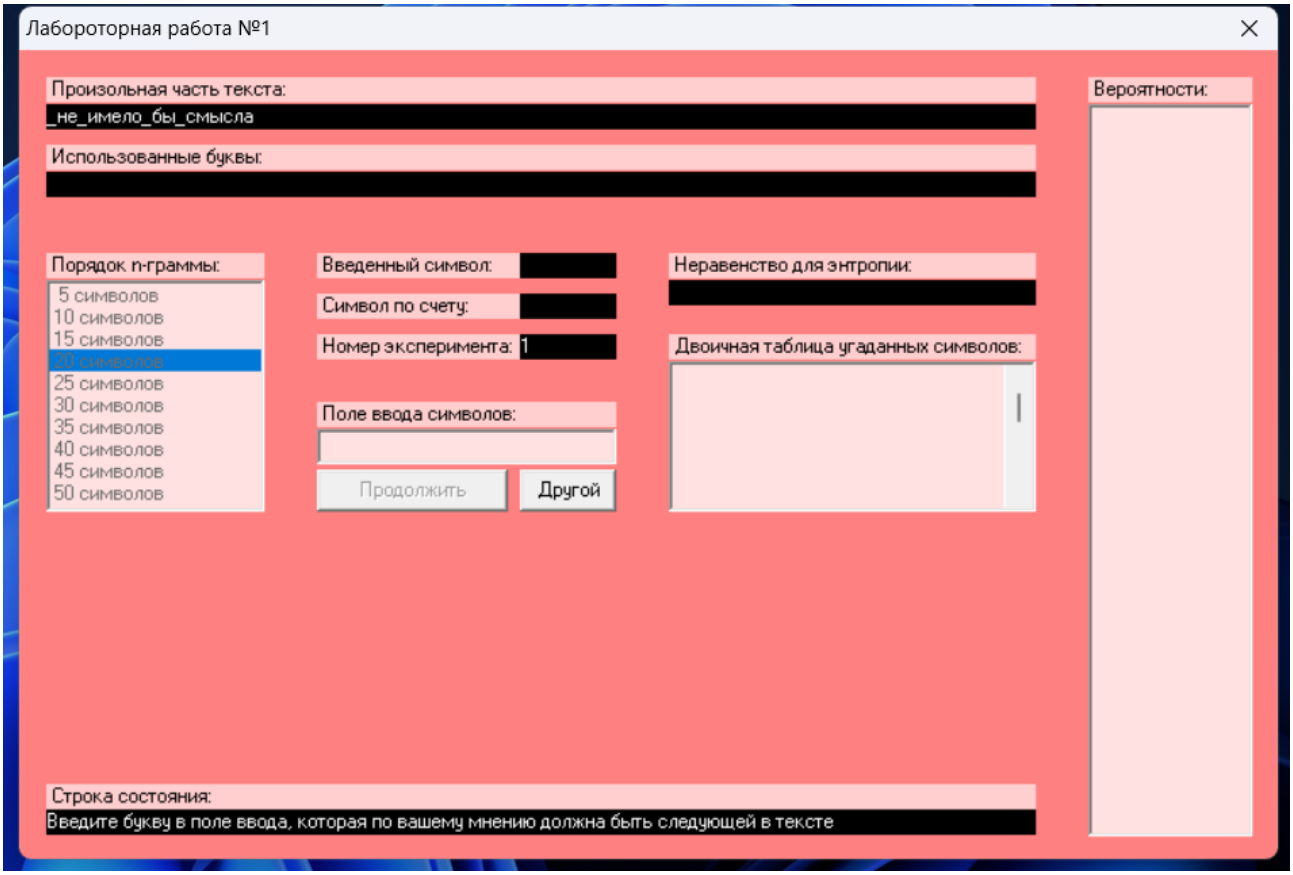
Виконання



$$H^{(10)}$$

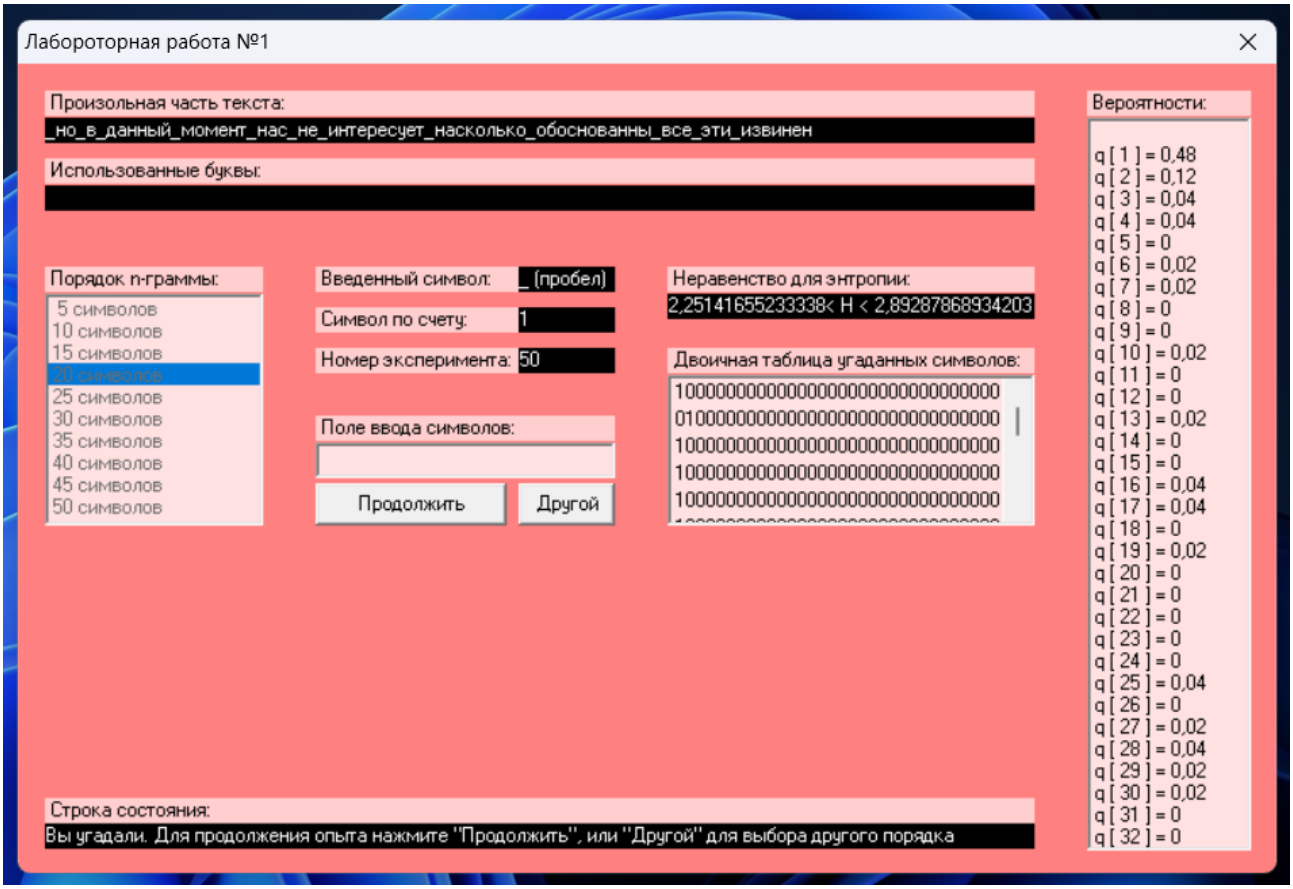
$$3{,}28539417202086 < H < 3{,}64836743955838$$

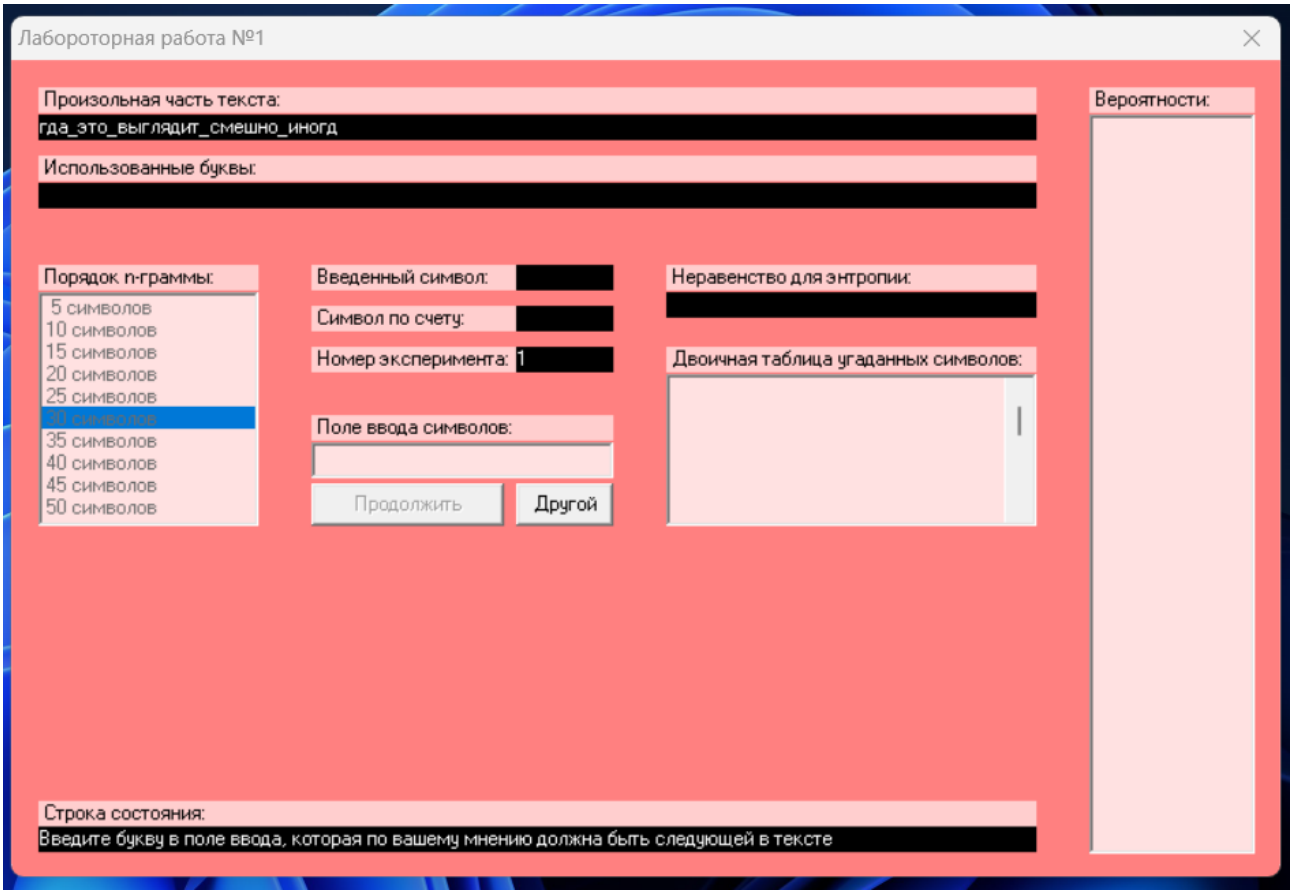




$$H^{(20)}$$

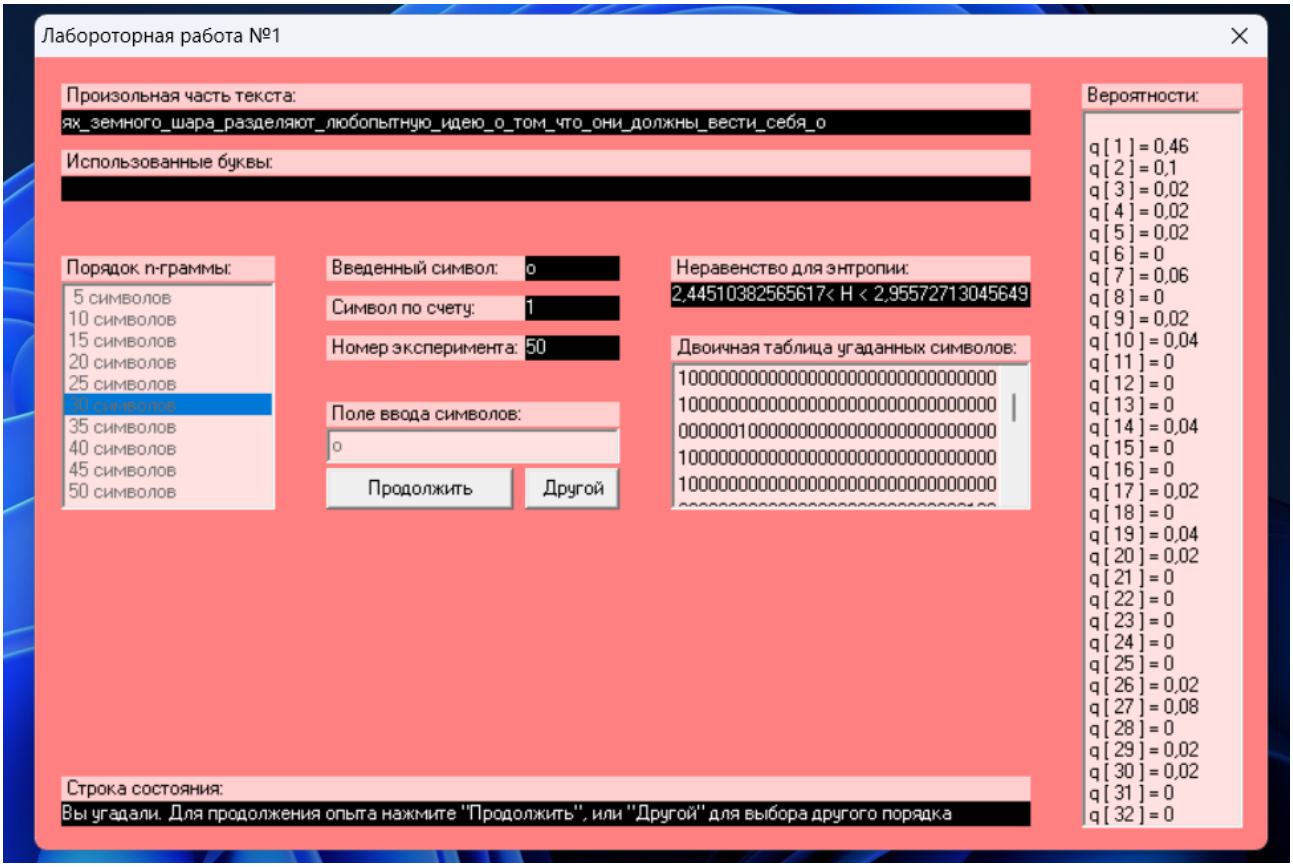
$$2{,}25141655233338 < H < 2{,}89287868934203$$





$$H^{(30)}$$

$$2{,}44510382565617 < H < 2{,}95572713045649$$



Завдання 3

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

Виконання

Для обрахунків використаємо цю формулу

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

$H_{(0)}$ можна знайти за наступною формулою, 32 це кількість літер у нашому алфавіті, це вказано у методичці, що у наданому тексті використовується лише 32 символи

$H_{(0)} = \log_2 32 = 5$

3,28539417202086 < H10 < 3,64836743955838

2,25141655233338 < H20 < 2,89287868934203

2,44510382565617 < H30 < 2,95572713045649

$H^{(10)}$

$$R = 1 - \frac{3,28539417202086}{5} \approx 0.34292116559582797$$

$$R = 1 - \frac{3,64836743955838}{5} \approx 0.27032651208832403$$

$H^{(20)}$

$$R = 1 - \frac{2,25141655233338}{5} \approx 0.5497166895333241$$

$$R = 1 - \frac{2,89287868934203}{5} \approx 0.421424262131594$$

$H^{(30)}$

$$R = 1 - \frac{2,44510382565617}{5} \approx 0.510979234868766$$

$$R = 1 - \frac{2,95572713045649}{5} \approx 0.40885457390870195$$

Висновки

У ході виконання роботи ми дослідили та оцінили ентропію для символів джерела відкритого тексту російською мовою. Було розроблено програму для підрахунку частот символів та біграм у тексті, а також обчислено ентропії H1 та H2 як з урахуванням пробілів, так і без них. Це дало можливість більш детально вивчити вплив пробілів на загальну ентропію тексту. Зокрема, результати показали, що включення пробілів впливає на розподіл частот окремих символів та біграм, проте не призводить до суттєвих змін значень H1 та H2 , що свідчить про низький внесок пробілів у загальну ентропію при обробці тексту значного обсягу.

Додатково, за допомогою програми CoolPinkProgram було обчислено ентропії H10, H20, H30 для тексту, що включає лише літери та пробіли, без урахування розділових знаків. Це дозволило оцінити надлишковість російської мови, яка, за нашими підрахунками, варіюється від 47% до 65% залежно від значення ентропії HHH. Високий рівень надлишковості свідчить про значну передбачуваність структури російської мови, що можна враховувати при проєктуванні систем стиснення тексту або криптографічного аналізу.

Результати також вказали на помітну різницю у значеннях H між текстом із пробілами та без них, а також між різними завданнями, що підтверджує залежність ентропії від обраного тексту та характеру символів, які він містить.