НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого текст

Виконали:
студентки групи ФБ-23
Гуз Вікторія
Шукалович Марія

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.

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 $H\ (10)$, $H\ (20)$, $H\ (30)$.

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

Хід роботи

Для початку ми залишили у тексті тільки російські літери(очистили від англійських, цифр, розділових знаків і подвійних пробілів):

```
def clean_text():
    with open('lab1.txt', 'r', encoding='utf-8') as file:
        content = file.read()
    content = content.lower()
    cleaned = re.sub(r'[^а-я\s]+', '', content)
    cleaned = re.sub(r'\s+', ' ', cleaned)
    cleaned = cleaned.strip()
    return cleaned
```

Також ми врахували, що моментами потрібно буде працювати з текстом без пробілів і очистили від них:

```
def remove_spaces(text):
    return text.replace(" ", "")
```

З цим етапом труднощів не виникало.

Далі слід було обчислити частоту кожної літери, щоб використати ці дані у формулі обчислення ентропії H1:

```
def letter_frequencies(letter_counts, total_count):
    frequencies = {}
    for letter, count in letter_counts.items():
        frequencies[letter] = count / total_count
    sorted_frequencies = dict(sorted(frequencies.items(), key=lambda item:
item[1], reverse=True))
    return sorted_frequencies
```

Також відсортували вивід частот у порядку спадання.
H1 обраховували за наступною формулою:

$$H_1 = -\sum_{i=1}^{n} p(i)\log_2 p(i)$$

p(i) – частота літери
n – кількість літер в алфавіті
Реалізація формули у коді:

```
def entropy_H1(letter_frequencies):
```

```
    entropy_value = 0
    for p_i in letter_frequencies.values():
        if p_i > 0:
            entropy_value -= p_i * math.log2(p_i)

    return entropy_value
```

Також обчислили частоту біграм(з/без перетином(-у)), щоб використати у формулі обчислення ентропії H2:

```
def bigram_frequencies(bigram_counts, total_bigrams):
    frequencies = {}
    for bigram, count in bigram_counts.items():
        frequencies[bigram] = count / total_bigrams
    return frequencies
```

Формула для обчислення H2:

$$H_2 = -\frac{1}{2}\sum_{i,j} p(i,j)\log_2 p(i,j)$$

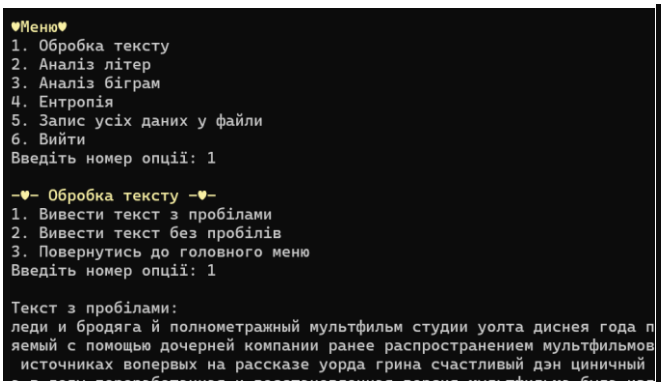
p(i,j) – частота біграми

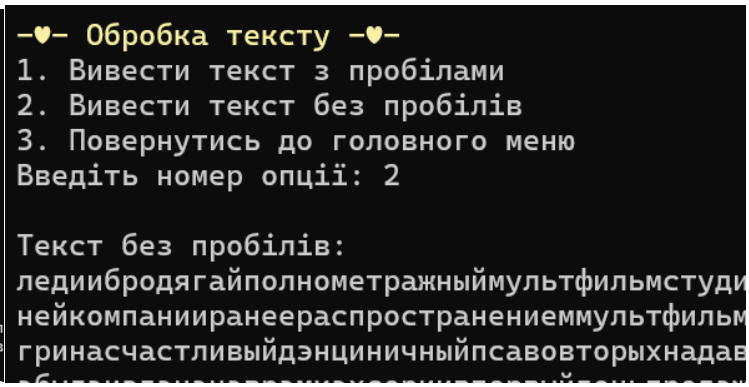
Реалізація формули у коді:

```
def entropy_H2(bigram_frequencies):
    entropy = 0
    for frequency in bigram_frequencies.values():
        if frequency > 0:
            entropy -= frequency * math.log2(frequency)
    return entropy / 2
```

Під час обчислення ентропії проблем не виникло. Усі необхідні кроки виконано правильно, а результати відповідали очікуванням. Формули для обчислення ентропії використовувалися без помилок, дані оброблялися належним чином, і процес обчислення був успішним

Приклад виконання:





```
-♥- Аналіз літер -♥-
1. Вивести кількість літер без пробілів
2. Вивести частоту літер без пробілів
3. Вивести кількість літер з пробілами
4. Вивести частоту літер з пробілами
5. Повернутись до головного меню
Введіть номер опції: 2

Частота літер:
Частота літери 'о': 0.11344491868895597
Частота літери 'е': 0.084954375903673O1
Частота літери 'а': 0.08020190203099696
Частота літери 'и': 0.065984402582096 47
Частота літери 'т': 0.065074871424579 21
Частота літери 'н': 0.064073345703419 94
Частота літери 'с': 0.052711149220389 47
Частота літери 'в': 0.046085804891680 74
Частота літери 'л': 0.045886194046042 41
Частота літери 'р': 0.043641873842474 06
Частота літери 'к': 0.033298560545544 0975
Частота літери 'д': 0.031738124456494 38
Частота літери 'м': 0.031325016793347 23
Частота літери 'у': 0.029566705605246 12
Частота літери 'п': 0.027624405289860984
```

```
-♥- Аналіз літер -♥-
1. Вивести кількість літер без пробілів
2. Вивести частоту літер без пробілів
3. Вивести кількість літер з пробілами
4. Вивести частоту літер з пробілами
5. Повернутись до головного меню
Введіть номер опції: 4

Частота літер:
Частота символу ' ': 0.16664713906114165
Частота символу 'о': 0.09453964754841761
Частота символу 'е': 0.07079697220860111
Частота символу 'а': 0.06683648451026936
Частота символу 'и': 0.05498829066913148
Частота символу 'т': 0.054230330276901446
Частота символу 'н': 0.053395705951869515
Частота символу 'с': 0.043926987006186635
Частота символу 'в': 0.03840573735515218
Частота символу 'л': 0.03823939108586505
Частота символу 'р': 0.036369080423358485
Частота символу 'к': 0.02774945069568903
Частота символу 'д': 0.026449056816653143
Частота символу 'м': 0.0261047923636937
Частота символу 'у': 0.024639498704668833
```

```
-♥- Аналіз біграм -♥-
--------------З пробілами---------------
1. Вивести кількість біграм з перетином
2. Вивести кількість біграм без перетину
3. Вивести частоту біграм з перетином
4. Вивести частоту біграм без перетину
--------------Без пробілів--------------
5. Вивести кількість біграм з перетином
6. Вивести кількість біграм без перетину
7. Вивести частоту біграм з перетином
8. Вивести частоту біграм без перетину
9. Повернутись до головного меню
Введіть номер опції: 7
      а       б       в       г       д       е
    ---------------------------------------------------
а | 0.00036 0.00139 0.00527 0.00125 0.00326 0.00229 0.
    ---------------------------------------------------
б | 0.00082 0.00001 0.00007 0.00000 0.00002 0.00246 0.
    ---------------------------------------------------
в | 0.00676 0.00024 0.00048 0.00030 0.00135 0.00635 0.
    ---------------------------------------------------
г | 0.00115 0.00003 0.00014 0.00002 0.00129 0.00042 0.
    ---------------------------------------------------
д | 0.00619 0.00007 0.00127 0.00002 0.00014 0.00496 0
```

```
-♥- Аналіз біграм -♥-
--------------З пробілами---------------
1. Вивести кількість біграм з перетином
2. Вивести кількість біграм без перетину
3. Вивести частоту біграм з перетином
4. Вивести частоту біграм без перетину
--------------Без пробілів--------------
5. Вивести кількість біграм з перетином
6. Вивести кількість біграм без перетину
7. Вивести частоту біграм з перетином
8. Вивести частоту біграм без перетину
9. Повернутись до головного меню
Введіть номер опції: 3
      а       б       в       г       д       е       ё
    -----------------------------------------------------------
а | 0.00006 0.00057 0.00278 0.00072 0.00184 0.00147 0.00000 0
    -----------------------------------------------------------
б | 0.00068 0.00000 0.00003 0.00000 0.00001 0.00204 0.00000 0
    -----------------------------------------------------------
в | 0.00556 0.00000 0.00003 0.00002 0.00086 0.00516 0.00000 0
    -----------------------------------------------------------
г | 0.00094 0.00000 0.00002 0.00000 0.00102 0.00033 0.00000 0
    -----------------------------------------------------------
```

```
-♥- Ентропія -♥-
1. H1 з пробілами
2. H1 без пробілів
3. H2 з пробілами, з перетинами
4. H2 без пробілів, з перетинами
5. H2 з проблілами, без перетинів
6. H2 без проблілами, без перетинів
7. Повернутись до головного меню
Введіть номер опції: 2

Ентропія H1: 4.45793

-♥- Ентропія -♥-
1. H1 з пробілами
2. H1 без пробілів
3. H2 з пробілами, з перетинами
4. H2 без пробілів, з перетинами
5. H2 з проблілами, без перетинів
6. H2 без проблілами, без перетинів
7. Повернутись до головного меню
Введіть номер опції: 3

Ентропія H2: 3.96032
```

```
♥Меню♥
1. Обробка тексту
2. Аналіз літер
3. Аналіз біграм
4. Ентропія
5. Запис усіх даних у файли
6. Вийти
Введіть номер опції: 5
Дані успішно записані у файли CSV.

♥Меню♥
1. Обробка тексту
2. Аналіз літер
3. Аналіз біграм
4. Ентропія
5. Запис усіх даних у файли
6. Вийти
Введіть номер опції: 6
 /}___/}❀
( • . •)
/ >    > Byeee
```

На наведених скриншотах лише частина даних для прикладу виконання, решта даних у прикріплених файлах:

- bigram_counts_no_overlap_with_spaces.csv – кількість і частота біграм без перетину, з пробілами;
- bigram_counts_no_overlap_without_spaces.csv – кількість і частота біграм без перетину, без пробілів;
- bigram_counts_with_spaces.csv – кількість і частота біграм з перетином, з пробілами;
- bigram_counts_without_spaces.csv - кількість і частота біграм з перетином, без пробілів;
- letter_counts_and_frequencies_with_spaces.csv – кількість і частота літер з пробілами;
- letter_counts_and_frequencies_without_spaces.csv – кількість і частота літер без пробілів;

До файлів не ввійшли біграми та літери з частотою 0, тому що неможливо взяти логарифм від 0.

Таблиця з отриманими значеннями ентропії:

	Без пробілів	З пробілами
H1	4.46598	4.37207
H2 з перетинами	4.14440	3.96426
H2 без перетинів	4.14597	3.96331

Таблиці з оцінками надлишковості російської мови:

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

Для тексту з пробілами $H_0 = \log_2 34 = 5.087$ (33 літери і пробіл)

Для тексту без пробілів $H_0 = \log_2 33 = 5.044$ (33 літери)

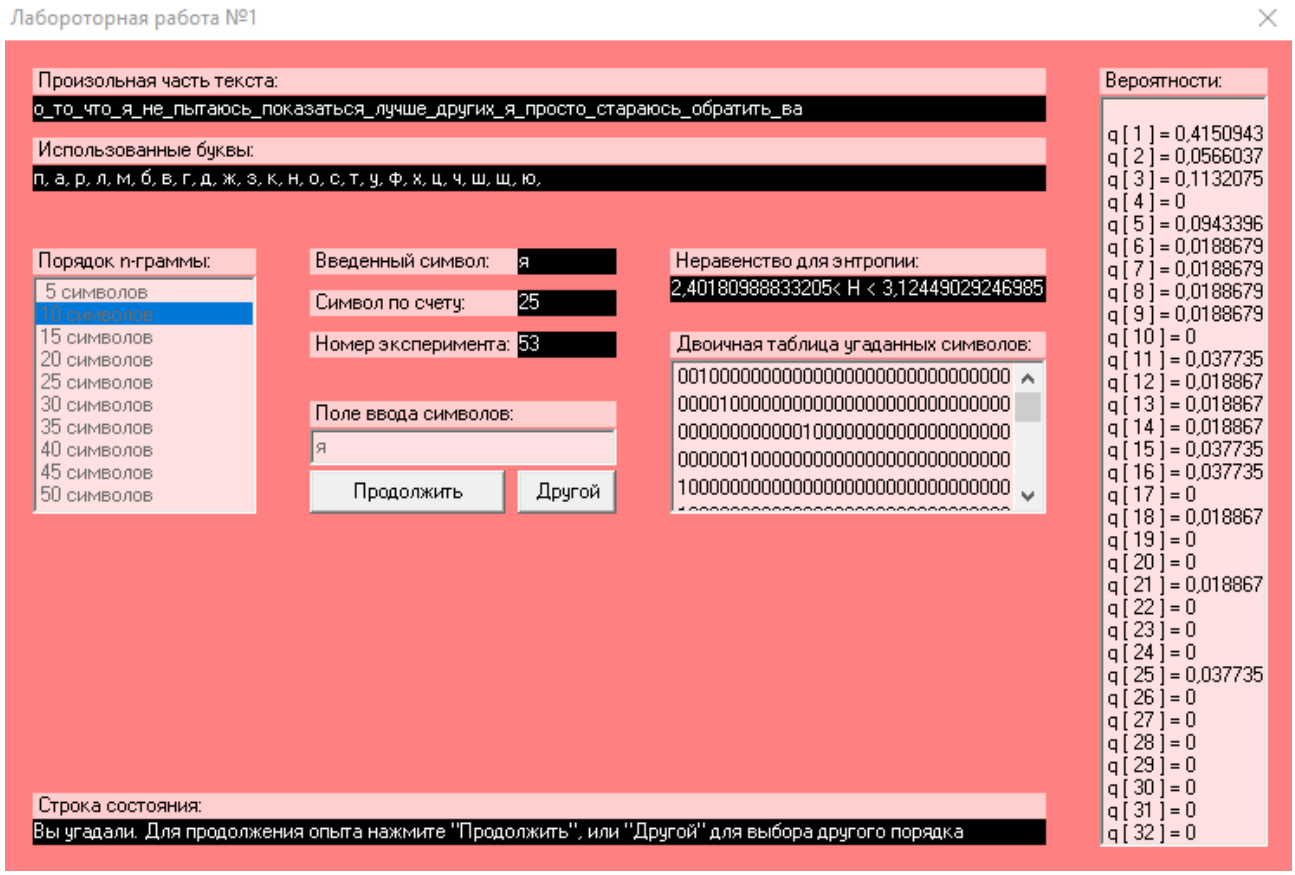
	Без пробілів	З пробілами
H1	0,11459	0,14054
H2 з перетинами	0,17835	0,2207
H2 без перетинів	0,17803	0,22089

Для тексту, який використовує CoolPinkProgram, $H_0 = \log_2 32 = 5$ (32 літери)

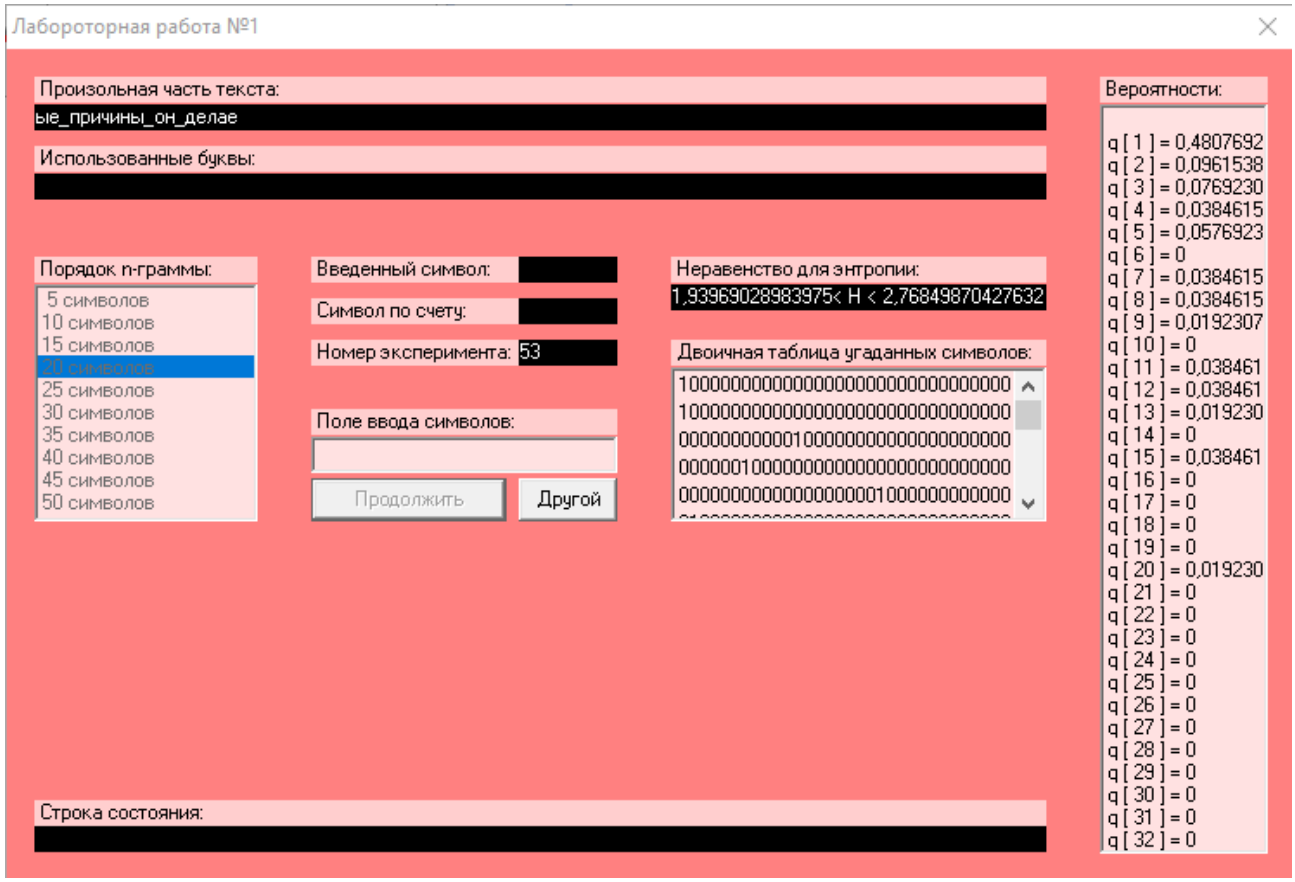
	Надлишковість
2,401809 <$H^{(10)}$<3,12449	0,51963>$R^{(10)}$>0,375101
1,93969<$H^{(20)}$<2,76849	0,612062>$R^{(20)}$>0,446302
1,18636<$H^{(30)}$<1,783405	0,76272>$R^{(30)}$>0,643319

CoolPinkProgram

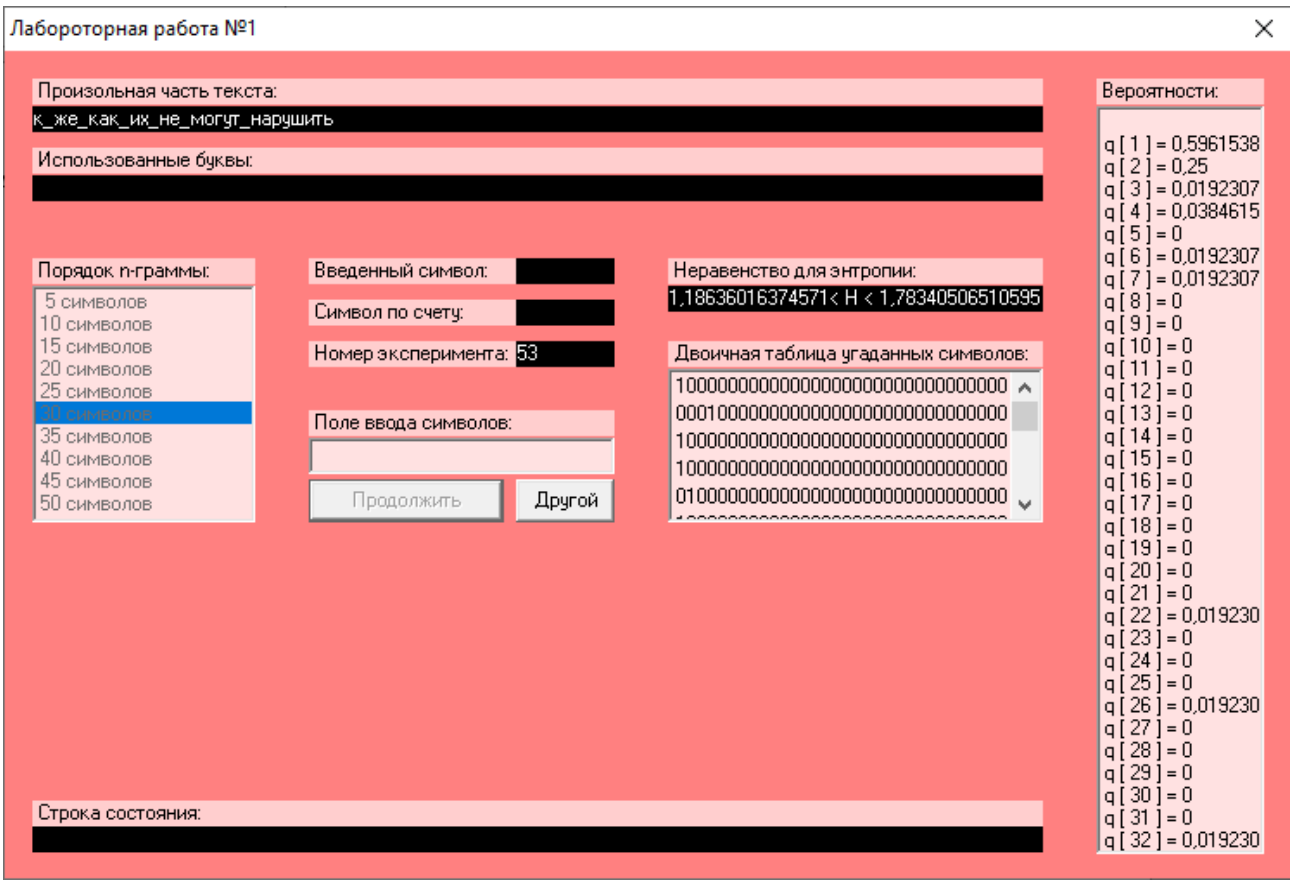
Для 10 символів:



Для 20 символів:



Для 30 символів:



Висновки: у ході виконання лабораторної роботи ми навчались визначати частоту літер та біграм в тексті, а також обраховувати значення його ентропії та надлишковості. Під час роботи з різними текстами ми помітили, що на ентропію суттєво впливають як довжина тексту, так і його структура (наявність повторюваних символів тощо). У текстах з високим рівнем впорядкованості (наприклад, з частими повтореннями тих самих слів чи символів) ентропія була нижчою, що вказує на меншу кількість інформації, необхідної для його передачі. Окрім цього, ми дослідили, що висока надлишковість свідчить про те, що текст містить багато передбачуваних елементів.