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

КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1 Експериментальна оцінка ентропії на символ джерела

відкритого тексту

Виконали студенти: ФБ-23 Лишиленко Ангеліна ФБ-23 Тіщенко Олександр

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 Н1 та Н2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення Н1 та Н2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення (10) Н, (20) Н, (30) Н.
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

1. Для того, аби виконати підрахунки, ми беремо російській текст. У нашому випадку це "Вечера на хуторе близ Диканьки", Гоголь Н. та починаємо писати кол:

Функція для підрахунку частот букв, в якій ми вибираємо у тексті лише букви російської мови

```
def let_freq(text):
    letter_counts = collections.Counter(text)
    total_letters = sum(letter_counts.values())
    letter_frequencies = {char: count / total_letters for char, count in letter_counts.items()}
    return letter_counts, letter_frequencies
```

Функція обчислює кількість і частоту біграм у тексті, з можливістю обрати варіант із перетином або без нього

```
def bi_freq(text, overlap=False):
    if overlap:
        bigrams = [text[i:i+2] for i in range(len(text)-1)]
    else:
        bigrams = [text[i:i+2] for i in range(0, len(text)-1, 2)]
    bigram_counts = collections.Counter(bigrams)
    total_bigrams = sum(bigram_counts.values())
    bigram_frequencies = {bigram: count / total_bigrams for bigram, count in bigram_counts.items()}
    return bigram_counts, bigram_frequencies
```

Функція обчислює ентропію Н1 на основі частот букв, використовуючи формулу

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

```
def calculate_H1(letter_frequencies):
    return -sum(frequency * math.log2(frequency) for frequency in letter_frequencies.values())
```

Функція обчислює ентропію Н2 на основі частот біграм, формула якої

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

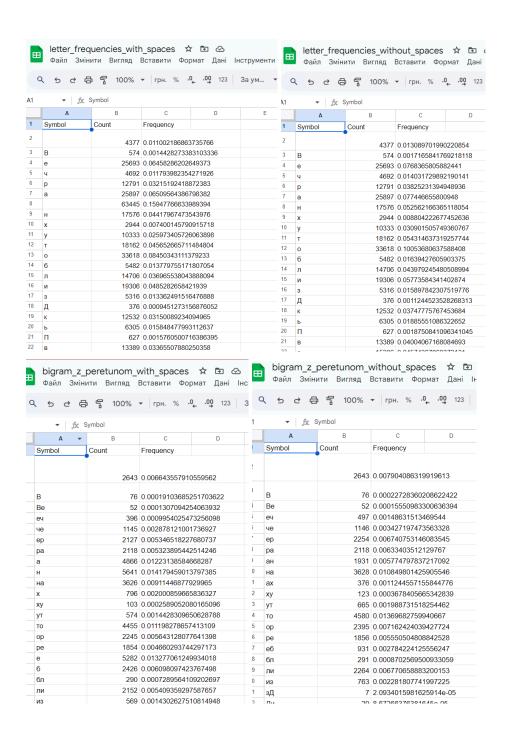
```
def calculate_H2(bigram_frequencies):
    return -sum(frequency * math.log2(frequency) for frequency in bigram_frequencies.values()) / 2
```

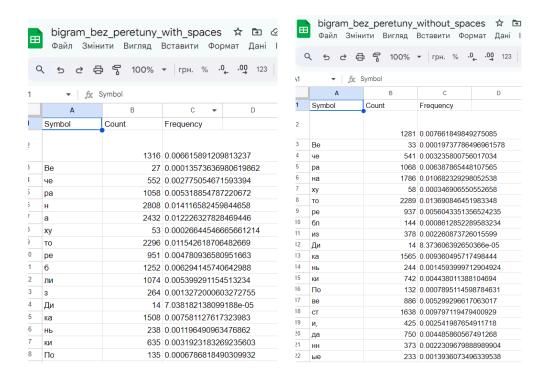
Далі в нас присутні функції для зберігання результату у cvs форматі, функція для обробки тексту, що повертає всі підрахунки, завантаження тексту та обробка з і без пробілів та виведення значень ентропії

В результаті ми отримуємо 6 таблиць та ось такий текст у консолі

```
ы bigram_bez_peretuny_with_spaces

Обробка тексту з пробілами...
Обробка тексту без пробілів...
Результат записано у CSV файли
Ентропія Н1 з пробілами: 4.724839741754606
Ентропія Н2 без перетину з пробілами: 4.196768910560601
Ентропія Н2 без перетину з пробілами: 4.197330532923184
Ентропія Н1 без пробілів: 4.868141008034572
Ентропія Н2 без перетину без пробілів: 4.411918771868499
Ентропія Н2 з перетином без пробілів: 4.4159115482839795
Ртеss any key to continue . . .
```



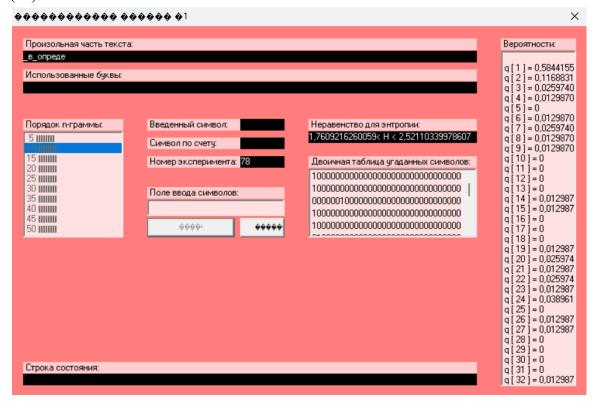


Отже наші значення ентропії

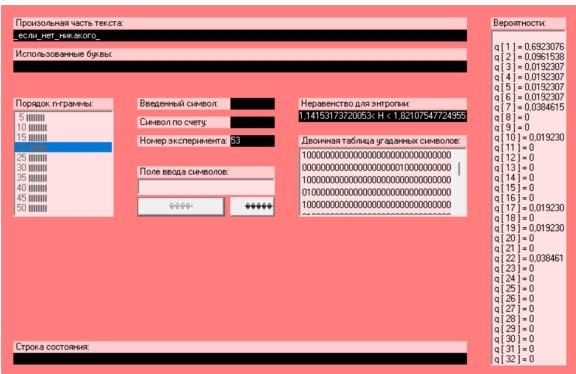
Ентропія	з пробілами	без пробілів
H1	4.7248397417546	4.8681410080345
Н2(з перетином)	4.1973305329231	4.4159115482839
Н2(без перетину)	4.1967689105606	4.4119187718684

2. Оцінка значень (10) H, (20) H, (30) H за допомогою програми CoolPinkProgram:

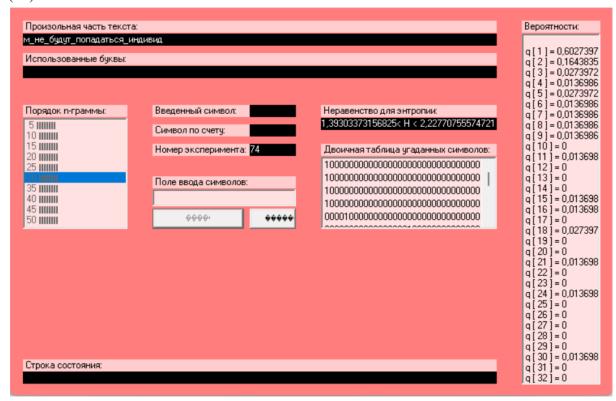
(10)H:



(20)H:



(30)H:



	Ентропія	Надлишковість
(10)H	1,7609216260059 <h<2,521103399 78607</h<2,521103399 	0.4957793200427859 <r<0,6478156747 9882</r<0,6478156747
(20)H	1,14153173720053 <h<1,82107547 724955</h<1,82107547 	0.63578490455009 <r<0.771693652559 894</r<0.771693652559
(30)H	1,39303373156825 <h<2,22770755 574721</h<2,22770755 	0.5544584888505579 <r<0.7213932536 8635</r<0.7213932536

Висновки:

У ході виконання лабораторної роботи ми закріпили знання з ентропії на символ джерела та надлишковості, також набули практичних навичок в експериментальній оцінці ентропії на символ джерела. Також дізнались про програму CoolPinkProgram яку в подальшому використали для знаходження ентропії.