### КРИПТОГРАФІЯ

#### КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

## **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

Мовою програмування для створення скриптів було обрано Python 3. Тому весь процес відбувався досить швидко, зручно і з малою кількістю дописування специфічних функцій.

Більшість труднощів, які виникали під час роботи, були пов'язані з:

- відсутністю досвіду роботи з GitHub;
- недостатньою компетентністю у математичних основах криптології (поняття ентропії, умовної ентропії), що фундаментально призупиняла написання коду;
- бажанням якомога краще оптимізувати набір створених функцій;
- несправність CoolPinkProgram.exe на основній і гостьовій ОС;
- поганим розумінням умови завдання.

Першу і другу проблеми було вирішено завдяки поглибленню у відповідні теми (здебільшого шляхом перегляду навчальних відео).

Приводом до появи третьої проблеми можна вважати відносну подібність між функціями або, навпаки, занадто розбіжні дії всередині однієї функції. Розглянемо пару таких інцидентів:

• count ngram() та find ngram() зведені до однієї ngram processing().

Цей випадок не зазначений в історії репозиторію, бо відбувся він через третю причину, яку ми обговоримо пізніше.

Так сталося, що мені потрібно було написати функцію для обрахування умовної ентропії. З'ясувалося, що наявної **entropy**() цілком достатньо, і треба лише дописати функції для підрахування "умовних" ймовірностей **conditional\_probability**(). Суть її полягала у тому, щоб підрахувати, скільки (n-1)-грам відповідає певній n-грамі, якщо n-1 їх символів однакові. Для цього було замало (хоча, скоріше, забагато) count\_ngram() і з'явилася потреба у функції пошуку n-грам без їх підрахунку. Так була створена find\_ngram(). Проте пізніше була помічена суттєва схожість між їх циклами, що вмотивувало об'єднати ці дві функції в одну **ngram\_processing**() з додатковим параметром, котрий визначав суть обробки.

Врешті-решт, хоча пошук і був визнаний непотрібним, він був залишений для повноти коду.

● Відділення від **read\_file**() функції **format\_text**().

Ця історія значно коротша за попередню.

Таке рішення було прийнято для більшої універсальності і повноти коду (типу, "а що як користувач не хоче досліджувати кастрований текст?").

Крім того, досить нелегко було вигадати самому, як представити частоту (ймовірність) появи біграм у вигляді матриці. Довелося поблукати мережею, і після недовгих пошуків таки-знайти зрозуміле і естетичне рішення.

Четверта проблема гальмувала увесь процес виконання лабораторної, адже спотворювало текст CoolPinkProgram.exe. З'ясувалося, що вона виникла через повну відсутність підтримки російської кирилиці у моїх системах. Була вирішена за пару хвилин.

Остання проблема має генетично-середовищне походження, тому не будемо на ній зупинятися (дякую за терпіння і регулярні відповіді на питання).

Розглянемо, отримані у ході виконання лабораторної роботи, дані. Варто зазначити, що обрахунки відбувалися для текстового файлу text.txt дуже значного обсягу ("Мертві душі" разом з "Війною та миром" фрагментами "Злочина і кари").

Усі результати обчислень були занесені у текстовий файл data.txt.

1. Частота літер (міститься у прикладеному файлі data.txt).

```
Frequencies (n = 1, with spaces) are:
    " " - 0.165
                     "o" - 0.095
                                      "e" - 0.072
                                                      "a" - 0.065
                                     "т" - 0.051
                                                      "c" - 0.042
    "и" - 0.057
                     "н" - 0.053
                                                      "K" - 0.034
                                     "p" - 0.035
    "в" - 0.038
                     "л" - 0.037
    "д" - 0.026
                     "M" - 0.025
                                     "y" - 0.025
                                                      "п" - 0.023
                                     "4" - 0.016
    "ь" - 0.017
                     "ы" - 0.016
                                                      "6" - 0.016
                     "г" - 0.014
                                     "з" - 0.014
    "я" - 0.016
                                                      "ж" - 0.009
    "й" - 0.009
                     "x" - 0.008
                                     "ш" - 0.008
                                                      "ю" - 0.005
                                     "э" - 0.002
    "ц" - 0.003
                     "щ" - 0.003
                                                      "φ" - 0.001
```

Знімок 1. Частоти літер у тексті з пробілами.

```
Frequencies (n = 1, without spaces) are:
                    "e" - 0.086
                                                     "и" - 0.068
    "o" - 0.114
                                    "a" - 0.078
    "н" - 0.063
                    "т" - 0.061
                                    "c" - 0.051
                                                     "B" - 0.045
                    "p" - 0.041
                                    "κ" - 0.041
    "л" - 0.045
                                                     "д" - 0.031
                                    "п" - 0.028
                    "v" - 0.03
                                                     "ь" - 0.021
    "м" - 0.03
    "ы" - 0.019
                    "4" - 0.019
                                    "6" - 0.019
                                                     "я" - 0.019
    "г" - 0.017
                    "a" - 0.017
                                    "ж" - 0.011
                                                     "x" - 0.01
                                                     "ц" - 0.003
    "й" - 0.01
                    "ш" - 0.01
                                    "ю" - 0.006
                    "э" - 0.002
                                    "d" - 0.001
    "щ" - 0.003
```

Знімок 2. Частоти літер у тексті без пробілів.

- 2. Частота біграм з пробілами та без них, з перетином біграм і без нього (міститься у прикладеному файлі data.txt).
- 3. Ентропія  $H_1$  та  $H_2$  (міститься у прикладеному файлі data.txt).

```
Entropy_n (n = 1, with spaces) is: 4.375304088464876
```

Знімок 3.  $H_1$  для тексту з пробілами.

```
Entropy_n (n = 1, without spaces) is:
    4.457392581982275
```

Знімок 4.  $H_1$  для тексту без пробілів.

```
Entropy_n (n = 2, with spaces) is:
3.7368895842238983
```

Знімок 5.  $H_2$  для тексту з пробілами.

# Entropy\_n (n = 2, without spaces) is: 3.817257841239711

Знімок 6.  $H_2$  для тексту без пробілів.

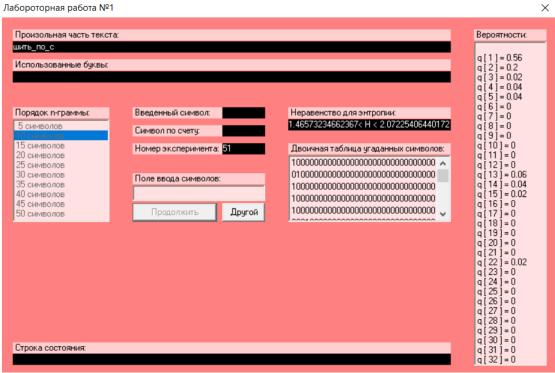
Entropy\_n (n = 2, with spaces, with step) is:
 3.710367916229299

3німок 7.  $H_2$  з перетином біграм для тексту з пробілами.

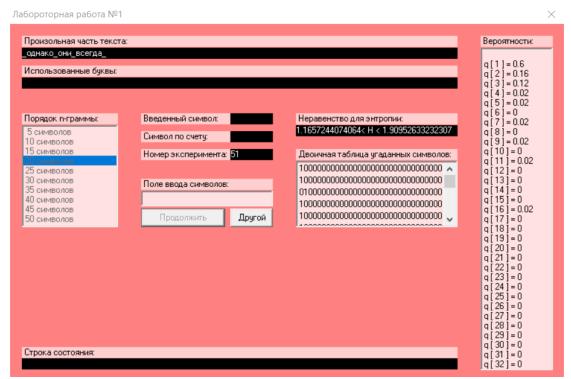
Entropy\_n (n = 2, without spaces, with step) is:
 3.8291077998811405

Знімок  $8.\,H_2$  з перетином біграм для тексту без пробілів.

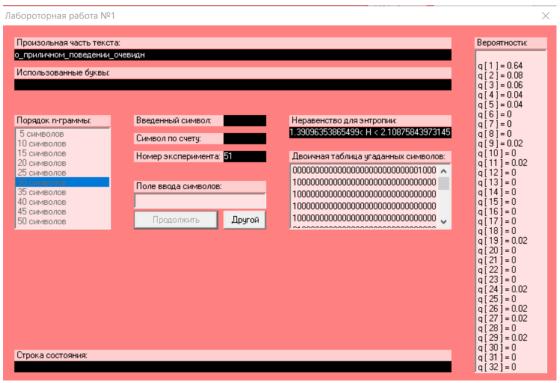
4. Оцінки для  $H^{(10)}$ ,  $H^{(20)}$  та  $H^{(30)}$ .



Знімок 9. Отримане значення  $H^{(10)}$ .



Знімок 10. Отримане значення  $H^{(20)}$ .



Знімок 11. Отримане значення  $H^{(30)}$ .

Обчислені значення в межах одного (n) можуть значно різнитися в залежності від кількості несприятливих випадків, коли рядок закінчувався на пробіл або на першій букві певного слова.

	найнижче значення	найвище значення
$H^{(10)}$	1.466	2.072
$H^{(20)}$	1.166	1.91
$H^{(30)}$	1.391	2.109

#### 5. Оцінка надлишковості R російської мови.

Здійснена у коді за формулою (міститься у прикладеному файлі data.txt):

$$R_n = 1 - \frac{H_n}{H_0} = 1 - \frac{H_n}{\log_2 m}$$

	з пробілами	без пробілів
$R_1$	0.125	0.1
$R_2$	0.253	0.229
R <sub>2</sub> (з кроком)	0.258	0.227

Здійснена у CoolPinkProgram.exe:

	найвище значення	найнижче значення
$R_{10}$	0.707	0.586
$R_{20}$	0.767	0.618
$R_{30}$	0.722	0.5782

Таким чином, значення  $R_n$  при малих n зростає з більшою швидкістю, ніж при n>10. Можна припустити, що надлишковість російського тексту фіксується приблизно на значенні 0.74.

#### Висновок

У ході виконання лабораторної роботи було обчислено частоти та ентропію літер і біграм у текстах російською мовою з пробілами та без них, з перетинами біграм та без них.

Крім того, було знайдено значення надлишковості російської мови як у коді, так і в програмі CoolPinkProgram.exe. Таким чином, було здійснено припущення, що надлишковість російської мови  $R_n$  при великих п прямує до 0.74.