



*Міністерство освіти і науки України Національний технічний університет України
«Київський політехнічний інститут ім. І. Сікорського» Фізико-технічний інститут*

КРИПТОГРАФІЯ

ЛАБОРАТОРНА РОБОТА №1

ЕКСПЕРИМЕНТАЛЬНА ОЦІНКА ЕНТРОПІЇ НА СИМВОЛ ДЖЕРЕЛА ВІДКРИТОГО ТЕКСТУ

Виконали:

Студенти групи ФБ-02

Лугінін Богдан

Хаустович Артем

Перевірила:

Байденко П. В.

Київ 2022 р.

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

1. Ознайомлення з теоретичним матеріалом, запропонованим у методичних рекомендаціях та вказівках до виконання комп'ютерного практикуму.
2. Реалізація програми для підрахунку частот букв і біграм у тексті, підрахунку H_1 та H_2 (згідно означення), підрахунку частот букв і біграм, значення H_1 та H_2 на основі довільно обраного тексту російською мовою достатньої довжини (у нашому випадку – файл *orwell.txt*), де імовірності замінити відповідними частотами. Також одержати значення H_1 і H_2 на тому ж тексті, в якому вилучено всі пробіли.
3. Оцінка за допомогою програми CoolPinkProgram значень $H(10)$, $H(20)$, $H(30)$.

Хід роботи

У ході роботи було обрано два твори Джорджа Орвелла “1984” та “Скотоферма” (які ми радимо прочитати). Так як умова роботи обмежує нас російською мовою, тому обидва твори, занесені до файлу *orwell.txt*, були запропоновані у російському перекладі.



При спробі “взяти” текстовий файл (це реалізує функція `inputtext()`), програма повернула помилку.

```
Traceback (most recent call last):
  File "C:\Users\vipar\OneDrive\Робочий стол\Lab_1\venv\Lab_1.py", line 145, in <module>
    text = re.sub(r'^[w\s]+|[[d]+', '', inputtext())
  File "C:\Users\vipar\OneDrive\Робочий стол\Lab_1\venv\Lab_1.py", line 6, in inputtext
    text = f.read()
  File "C:\Users\vipar\AppData\Local\Programs\Python\Python310\lib\codecs.py", line 322, in decode
    (result, consumed) = self._buffer_decode(data, self.errors, final)
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xc4 in position 0: invalid continuation byte
```

Використовуючи літературу у вільному доступі, було запропоноване таке вирішення проблеми. Заміна параметру `utf-8` на `latin-1`. У результаті отримали таке.

Було: <pre>def inputtext(): f = open("text.txt", "r", encoding='utf-8') text = f.read() text = text.replace("\n", "") text = text.lower() f.close() return text</pre>	Стало: <pre>def inputtext(): f = open("orwell.txt", "r", encoding='latin-1') text = f.read() text = text.replace("\n", "") text = text.lower() f.close() return text</pre>
---	--

В результаті отримано програму (файл Lab_1.py.

Робота з програмою CoolPinkProgram

Пошук Н (10).

Лабораторная работа №1

Произвольная часть текста:
_странц_где_восхищаются_людьми_которые_убегают_с_поля_битвы_или_где_человек

Использованные буквы:
я, ч, с, _, м, и, т, ь, б, ю, э, ж,

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: д

Символ по счету: 13

Номер эксперимента: 81

Неравенство для энтропии:
 $4,68600552963363 < H < 4,35004101613061$

Двоичная таблица угаданных символов:
00000000000000000000000000000000
00000000000000000000000000000000
00000000000000000000000000000000
00000000000000000000000000000001
00000000000000000000000000000000
00000000000000000000000000000000

Поле ввода символов:
д

Продолжить Другой

Вероятности:
q[1] = 0,0987654
q[2] = 0
q[3] = 0,0123456
q[4] = 0,0123456
q[5] = 0,0493827
q[6] = 0,0617283
q[7] = 0,0493827
q[8] = 0,0370370
q[9] = 0,0123456
q[10] = 0,024691
q[11] = 0
q[12] = 0
q[13] = 0,024691
q[14] = 0,012345
q[15] = 0,086419
q[16] = 0,061728
q[17] = 0,037037
q[18] = 0,037037
q[19] = 0,111111
q[20] = 0,024691
q[21] = 0,024691
q[22] = 0,037037
q[23] = 0
q[24] = 0,037037
q[25] = 0,024691
q[26] = 0,049382
q[27] = 0,049382
q[28] = 0
q[29] = 0,012345
q[30] = 0
q[31] = 0
q[32] = 0,012345

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$$4.686 < H^{10} < 4.350$$

$$0,06 < R < 0,13$$

Пошук Н (20).

[illegible]

$$5.149 < H^{20} < 4.566$$

$$0,02 < R < 0,08$$

Пошук Н (30).

The screenshot shows a software interface for calculating entropy. At the top, a title bar reads "Лабораторная работа №1" (Laboratory work №1) with a close button. The main window is divided into several sections:

- Top Section:** A text input field contains "Произвольная часть текста:" (Arbitrary part of the text:). Below it, a black bar displays the text "век_может_нарушить_по_своему_". Below that, a label "Использованные буквы:" (Used letters:) is followed by another black bar.
- Left Panel:** A list titled "Порядок n-граммы:" (Order of n-grams:) shows options from 5 to 50 symbols. The option "10 символов" (10 symbols) is selected and highlighted in blue.
- Center Section:** A label "Введенный символ:" (Entered symbol:) is followed by a black bar. Below it, a label "Символ по счету:" (Symbol by number:) is followed by another black bar. A label "Номер эксперимента:" (Experiment number:) is followed by a black bar containing the number "85". Below these is a label "Поле ввода символов:" (Symbol input field:) followed by an empty text box. At the bottom of this section are two buttons: "Продолжить" (Continue) and "Другой" (Other).
- Right Section:** A label "Вероятности:" (Probabilities:) is followed by a list of probabilities for symbols 1 through 32. The list shows values like $q[1] = 0,0357142$, $q[2] = 0$, $q[3] = 0,0119047$, etc., up to $q[32] = 0,178571$.
- Bottom Section:** A label "Строка состояния:" (Status line:) is followed by a black bar.
- Entropy Calculation:** A label "Неравенство для энтропии:" (Entropy inequality:) is followed by a black bar displaying the inequality $5,14089601012869 < H < 4,12188298381944$.
- Symbol Table:** A label "Двоичная таблица угаданных символов:" (Binary table of guessed symbols:) is followed by a table with two columns. The first column contains binary strings of length 10, and the second column contains symbols. The symbols are mostly "0" and "1", with some "2" and "3" at the bottom.

$$5.140 < H^{30} < 4.121$$

$$0,02 < R < 0,76$$

Висновки

У ході виконання цього комп'ютерного практикуму було досліджено ентропію на символ джерела та його надлишковості, вивчено та порівняно різні моделі джерел відкритого тексту для наближеного визначення ентропії, набуто практичних навичок щодо оцінки ентропії на символ джерела.

Найчастіші літери в алфавіті:

а, е, к, о, р, с, т, у, я, “ ”.