



*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).

[illegible]

Пошук Н (20).

Лабораторная работа №1

Произвольная часть текста:
шестьсот_добро_и_зло_нельзя_словами_если_нет_никакого_закона_человеческой_пр

Использованные буквы:

Порядок итерации:

- 1 символ
- 5 символов
- 10 символов
- 15 символов
- 20 символов
- 25 символов
- 30 символов
- 35 символов
- 40 символов
- 45 символов
- 50 символов

Введенный символ: a

Символ по счету: 1

Интер-эксперимента: 89

Неравенство для энтропии:
1.953203215841001 * N < 2.69702384056762

Двоичная таблица значений символов:

0100000000000000000000000000000000	▲
0100000000000000000000000000000000	
0100000000000000000000000000000000	
0100000000000000000000000000000000	
0001000000000000000000000000000000	▼

Поле ввода символов:
a

Продолжить Другой

Вероятности:

- q[1] = 0.5306122
- q[2] = 0.0016309
- q[3] = 0.0488163
- q[4] = 0.0486163
- q[5] = 0.0234081
- q[6] = 0
- q[7] = 0.0006163
- q[8] = 0
- q[9] = 0
- q[10] = 0
- q[11] = 0.049816
- q[12] = 0.001324
- q[13] = 0.0488163
- q[14] = 0.023408
- q[15] = 0
- q[16] = 0
- q[17] = 0.023408
- q[18] = 0.023408
- q[19] = 0
- q[20] = 0.023408
- q[21] = 0
- q[22] = 0
- q[23] = 0
- q[24] = 0
- q[25] = 0.023408
- q[26] = 0
- q[27] = 0
- q[28] = 0.023408
- q[29] = 0
- q[30] = 0
- q[31] = 0
- q[32] = 0

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

Пошук Н (30).

Лабораторная работа №1

Произвольная часть текста:
оцениваться различными биологическими законами, которые он не может нарушить по...

Использованные буквы:

Порядок отгадывания:

- 5 символов
- 10 символов
- 15 символов
- 20 символов
- 25 символов
- 30 символов
- 35 символов
- 40 символов
- 45 символов
- 50 символов

Введенный символ: **е**

Символ по счету: **1**

Номер эксперимента: **50**

Неравенство для энтропии: $H \leq 2.32669214332165$

Двоичная таблица угаданных символов:

1000000000000000000000000000000000
1000000000000000000000000000000000
0100000000000000000000000000000000
0000100000000000000000000000000000
1000000000000000000000000000000000

Поле ввода символа: **е**

Правильный Другой

Вероятности:

q[1] = 0.6
q[2] = 0.14
q[3] = 0.02
q[4] = 0.02
q[5] = 0.04
q[6] = 0.02
q[7] = 0
q[8] = 0
q[9] = 0.02
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0.02
q[17] = 0.02
q[18] = 0
q[19] = 0.02
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0.02
q[26] = 0.02
q[27] = 0.02
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0.02

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

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, “ ”.