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 H1та H2 (згідно означення), підрахунку частот букв і біграм, значення H1 та H2 на основі довільно обраного тексту російською мовою достатньої довжини (у нашому випадку – файл *orwell.txt*), де імовірності замінити відповідними частотами. Також одержати значення H1 і H2 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\Рабочий стол\Lab_1\venv\Lab_1.py", line 145, in <module>
    text = re.sub(r'[^\w\s]+|[\d]+', '', inputtext())
  File "C:\Users\vipar\OneDrive\Рабочий 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.

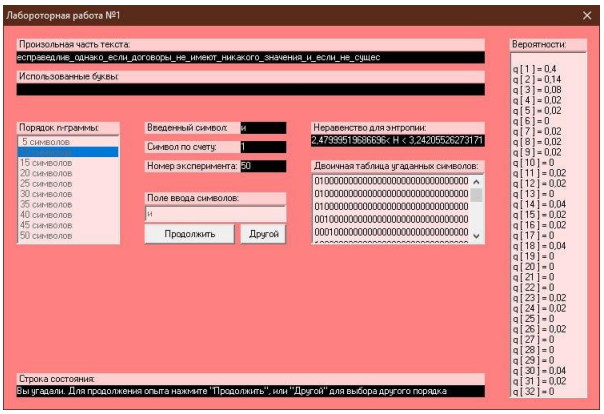
Було:	**Стало:**

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

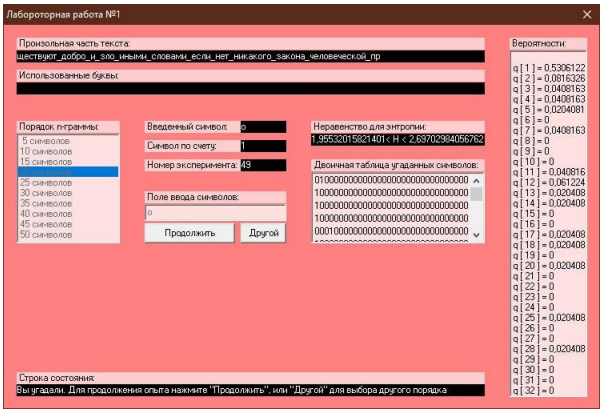
В результаті отримано програму (файл Lab_1.py.

Робота з програмою CoolPinkProgram

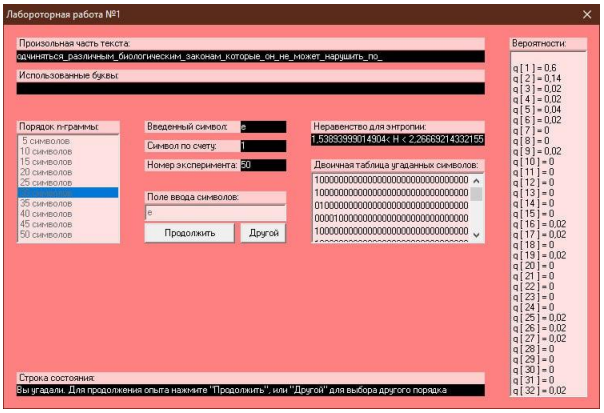
Пошук H (10).



Пошук H (20).



Пошук H (30).



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, “ ”.