**Міністерство освіти і науки України**

**Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"**

**Факультет інформатики та обчислювальної техніки**

**Кафедра інформатики та програмної інженерії**

**Звіт**

з лабораторної роботи №2 з дисципліни

«Обробка та аналіз текстових даних на Python»

„**Попередня обробка тексту за допомогою NLTK**”

**Виконав(ла)**

(шифр, прізвище, ім'я, по батькові)

*ІП-11 Головня Олександр Ростиславович*

**Перевірив**

(прізвище, ім'я, по батькові)

Київ 2024

**Мета роботи:** Ознайомитись з основними операціями з попередньої

обробки тексту та їх реалізацією у бібліотеці NLTK.

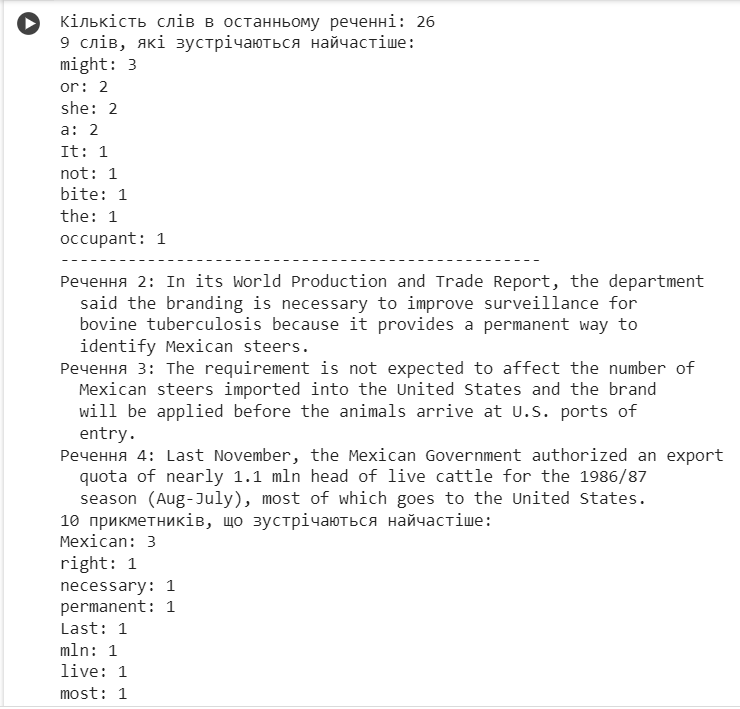
**Завдання до лабораторної роботи:**

Створити програму, яка виконує завдання відповідно до варіанту,

використовуючи бібліотеку NLTK.

1.Зчитати файл text2. а) Порахувати кількість слів в останньому реченні (не враховуючи знаки пунктуації та інші спеціальні символи); б) позначити частини мови; в) знайти 9 слів, які зустрічаються найчастіше.  
2.Використати корпус Reuters, четвертий текст категорії livestock. а) Вивести речення з другого до п’ятого; б) Вивести 10 прикметників, що зустрічаються найчастіше.

Результат виконання коду



**Код програми:**

import nltk

# nltk.download('punkt')

# nltk.download('averaged\_perceptron\_tagger')

#nltk.download('reuters')

from nltk.corpus import reuters

from nltk.probability import FreqDist

from nltk.tokenize import sent\_tokenize, word\_tokenize

from nltk.tag import pos\_tag

# 1.Зчитати файл text2. а) Порахувати кількість слів в останньому реченні

# (не враховуючи знаки пунктуації та інші спеціальні символи);

# б) позначити частини мови; в) знайти 9 слів, які зустрічаються найчастіше.

# 2.Використати корпус Reuters, четвертий текст категорії livestock.

# а) Вивести речення з другого до п’ятого;

# б) Вивести 10 прикметників, що зустрічаються найчастіше.

with open('text2.txt', 'r' ) as file:

  text = file.read()

sentences = sent\_tokenize(text)

words = word\_tokenize(text)

last\_sentence = sentences[-1]

words\_in\_last\_sentence = [word.strip('.,?!') for word in word\_tokenize(last\_sentence) if word.isalnum()]

num\_words\_in\_last\_sentence = len(words\_in\_last\_sentence)

print(f"Кількість слів в останньому реченні: {num\_words\_in\_last\_sentence}")

# pos\_tags = pos\_tag(words)

# for word, pos\_tag in pos\_tags:

#   print(f"{word}: {pos\_tag}")

fdist = FreqDist(words\_in\_last\_sentence)

most\_common\_words = fdist.most\_common(9)

print("9 слів, які зустрічаються найчастіше:")

for word, freq in most\_common\_words:

    print(f"{word}: {freq}")

print("--------------------------------------------------")

# Використати корпус Reuters, четвертий текст категорії livestock

livestock\_fileids = reuters.fileids(categories='livestock')

fourth\_livestock\_fileid = livestock\_fileids[3]

livestock\_text = reuters.raw(fileids=fourth\_livestock\_fileid)

# print(livestock\_text)

livestock\_sentences = sent\_tokenize(livestock\_text)

livestock\_words = word\_tokenize(livestock\_text)

for i in range(1, 4):  # ВСЬОГО 4!

    print(f"Речення {i+1}: {livestock\_sentences[i]}")

livestock\_pos\_tags = pos\_tag(livestock\_words)

adjectives = [word for word, tag in livestock\_pos\_tags if tag.startswith('JJ')]

fdist\_adj = FreqDist(adjectives)

most\_common\_adjectives = fdist\_adj.most\_common(10)

print("10 прикметників, що зустрічаються найчастіше:")

for adj, freq in most\_common\_adjectives:

    print(f"{adj}: {freq}")

**Висновок:** Під час розробки лабораторної роботи, я ознайомився з основними операціями з попередньої обробки тексту та їх реалізацією у бібліотеці NLTK.