Національний технічний університет України «КПІ ім. Ігоря Сікорського»
Факультет Інформатики та Обчислювальної Техніки
Кафедра інформаційних систем та технологій

Лабораторна робота № 2
з дисципліни «Обробка та аналіз текстових даних на Python»
На тему:
«Попередня обробка тексту за допомогою NLTK»
Варіант №3

Виконала:
студентка групи ІС-12.
Мельникова К.О.
Перевірила:
Тимофєєва Ю. С.

Київ – 2024

Мета роботи: Ознайомитись з основними операціями з попередньої обробки тексту та їх реалізацією у бібліотеці NLTK.

Завдання до лабораторної роботи

1.Зчитати файл text3. а) Порахувати кількість слів в тексті (не враховуючи знаки пунктуації та інші спеціальні символи); б) видалити стоп-слова; в) позначити частини мови та вивести третє речення з анотованими словами.

Спершу зчитаємо файл:

Лабораторна робота №2

Зчитаємо файл:

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.tag import pos_tag
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('averaged_perceptron_tagger')
nltk.download('brown')
nltk.download('maxent_ne_chunker')
nltk.download('words')

# Зчитати файл text3
with open('data.txt', 'r', encoding='utf-8') as file:
    text = file.read()
```

✓ 0.0s Python

Виконаємо перше завдання:

```
# 1. а) Порахувати кількість слів в тексті (не враховуючи знаки пунктуації та інші спеціальні символи)
words = word_tokenize(text.lower())  # токенізація тексту та переведення у нижній регістр
words = [word for word in words if word.isalpha()]  # видалення всіх слів, що не містять лише букви
num_words = len(words)
print("Кількість слів в тексті (без знаків пунктуації та спеціальних символів):", num_words)

# б) Видалити стоп-слова
stop_words = set(stopwords.words('english'))
filtered_words = [word for word in words if word not in stop_words]
filtered_words_line = " ".join(filtered_words)
print("Текст без стоп-слів: " + filtered_words_line)

# в) Позначити частини мови та вивести третє речення з анотованими словами
sentences = sent_tokenize(text)
tagged_words = pos_tag(filtered_words)
third_sentence = sentences[2]
print("\nТретє речення з анотованими словами:")
print(nltk.ne_chunk(tagged_words))
```

[18] ✓ 0.0s Python

```
Кількість слів в тексті (без знаків пунктуації та спеціальних символів): 168
Текст без стоп-слів: weeks marriage days still sharing rooms holmes baker street came home afternoon stroll find letter tabl

Третє речення з анотованими словами:
(S
  weeks/NNS
  marriage/NN
  days/NNS
  still/RB
  sharing/VBG
  rooms/NNS
  holmes/RB
  baker/VBP
  street/NN
  came/VBD
  home/RB
  afternoon/NN
  stroll/NN
  find/VBP
  letter/NN
  table/NN
  waiting/VBG
  remained/VBD
  day/NN
  weather/NN
```

2.Використати корпус Brown, третій текст категорії editorial. а) Порахувати загальну кількість речень; б) Видалити всі іменники.

Друге завдання:

```python
# 2. а) Порахувати загальну кількість речень у корпусі Brown, категорії editorial
editorial_sentences = nltk.corpus.brown.sents(categories='editorial')
num_editorial_sentences = len(editorial_sentences)
print("\nЗагальна кількість речень у категорії editorial корпусу Brown:", num_editorial_sentences)

# б) Видалити всі іменники
editorial_words = nltk.corpus.brown.words(categories='editorial')
print("З іменниками: ")
print([word for word, _ in pos_tag(editorial_words)])
editorial_without_nouns = [word for word, pos in pos_tag(editorial_words) if pos != 'NN']
print("Без іменників: ")
print(editorial_without_nouns[:200])
```

21] ✓ 2.2s Python

```
Загальна кількість речень у категорії editorial корпусу Brown: 2997
З іменниками:
['Assembly', 'session', 'brought', 'much', 'good', 'The', 'General', 'Assembly', ',', 'which', 'adjourns', 'today', ',', 'ha
Без іменників:
['Assembly', 'brought', 'much', 'good', 'The', 'General', 'Assembly', ',', 'which', 'adjourns', ',', 'has', 'performed', 'in
```