Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем і технологій

Лабораторна робота №2

З дисципліни «Обробка та аналіз текстових даних на Python»

Варіант 13

Виконав студент гр. ІС-11

Бондарчук О. А.

Перевірила ст. в. каф. ІСТ

Тимофєєва Ю. С.

Київ 2024

Завдання: 1.Зчитати файл text3. а) Порахувати кількість слів в третьому реченні (не враховуючи знаки пунктуації та інші спеціальні символи); б) видалити стоп-слова; в) знайти 10 слів, які зустрічаються найчастіше.  
2.Використати корпус **Reuters**, третій текст категорії **cocoa**. а) Вивести речення з другого до передостаннього; б) Вивести 8 дієслів, що зустрічаються найчастіше.

Початковий текст:

It was a few weeks before my own marriage, during the days when I was  
still sharing rooms with Holmes in Baker Street, that he came home from  
an afternoon stroll to find a letter on the table waiting for him.  
I had remained in-doors all day, for the weather had taken a sudden  
turn to rain, with high autumnal winds, and the jezail bullet which I  
had brought back in one of my limbs as a relic of my Afghan campaign,  
throbbed with dull persistency. With my body in one easy-chair and my  
legs upon another, I had surrounded myself with a cloud of newspapers,  
until at last, saturated with the news of the day, I tossed them all  
aside and lay listless, watching the huge crest and monogram upon the  
envelope upon the table, and wondering lazily who my friend’s noble  
correspondent could be.  
  
“Here is a very fashionable epistle,” I remarked, as he entered. “Your  
morning letters, if I remember right, were from a fish-monger and a  
tide-waiter.”

Лістинг програми

from nltk.corpus import stopwords  
from nltk.tokenize import word\_tokenize, sent\_tokenize  
from collections import Counter  
from nltk.tag import pos\_tag  
from nltk.corpus import reuters  
  
def read\_file\_to\_string(fileName: str):  
 try:  
 with open(fileName, mode='r') as file:  
 content = file.read()  
 return content  
 except FileNotFoundError:  
 print(f"Error: File '{fileName}' not found.")  
 return None  
 finally:  
 file.close()  
  
  
filename: str = "text3.txt"  
  
text = read\_file\_to\_string(filename)  
  
# Розбиваємо текст на речення  
sentences = sent\_tokenize(text)  
  
# Вибираємо третє речення (індекс 2, оскільки індекси починаються з 0)  
third\_sentence = sentences[2]  
  
# а) Пораховуємо кількість слів у третьому реченні (без пунктуації та інших спеціальних символів)  
words = word\_tokenize(third\_sentence)  
cleaned\_words = [word.lower() for word in words if word.isalnum()] # Видаляємо всі знаки пунктуації  
word\_count = len(cleaned\_words)  
print("Кількість слів у третьому реченні:", word\_count)  
  
# б) Видаляємо стоп-слова  
stop\_words = set(stopwords.words('english'))  
filtered\_words = [word for word in cleaned\_words if word not in stop\_words]  
  
# в) Знаходимо 10 слів, які зустрічаються найчастіше  
word\_freq = Counter(filtered\_words)  
top\_10\_words = word\_freq.most\_common(10)  
print("Топ 10 слів, які зустрічаються найчастіше:")  
for word, freq in top\_10\_words:  
 print(f"{word}: {freq}")  
  
# Завантажуємо корпус Reuters та вибираємо тексти категорії cocoa  
cocoa\_docs = reuters.fileids(categories='cocoa')  
  
# Вибираємо третій текст категорії cocoa (індекс 2, оскільки індекси починаються з 0)  
third\_cocoa\_text = reuters.raw(fileids=cocoa\_docs[2])  
print("\nthird cocoa text:\n" + third\_cocoa\_text)  
  
# Розбиваємо текст на речення  
sentences = sent\_tokenize(third\_cocoa\_text)  
  
# а) Виводимо речення з другого до передостаннього  
print("Речення з другого до передостаннього:")  
for sentence in sentences[1:-1]:  
 print(sentence)  
  
# б) Вибираємо дієслова з тексту та обчислюємо їх частоту  
verbs = []  
for sentence in sentences:  
 words = word\_tokenize(sentence)  
 tagged\_words = pos\_tag(words)  
 for word, tag in tagged\_words:  
 if tag.startswith('VB'): # VB - дієслово  
 verbs.append(word.lower())  
  
# Знаходимо 8 дієслів, які зустрічаються найчастіше  
verb\_freq = Counter(verbs)  
top\_8\_verbs = verb\_freq.most\_common(8)  
print("\n8 найчастіше зустрічаючихся дієслів:")  
for verb, freq in top\_8\_verbs:  
 print(f"{verb}: {freq}")

Результати виконання програми:

Завдання 1:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

Завдання 2:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис  
Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

Зображення, що містить текст, знімок екрана, Шрифт, дизайн

Автоматично згенерований опис

Посилання на Github:  
<https://github.com/SashaBondarchuk/text-data-analysis-labs/tree/main>