МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

**ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ №16**

з дисципліни «Програмування мовою Python»

на тему «обка природної мови з використанням Python бібліотек»

Виконав студент групи КН-32

Ярош Я.О.

Перевірила Парфененко Ю.В.

Варіант 7

Суми 2024

**Завдання 1**

1. **Постановка задачі**

Імпортувати бібліотеку NLTK та тексти із електронного архіву текстів Project Gutenberg, для виконання завдань взяти текст, заданий варіантом. Визначити кількість слів у текст. Визначити 10 найбільш вживаних слів у тексті, побудувати на основі цих даних стовпчасту діаграму. Виконати видалення з тексту стоп-слів та пунктуації, після чого знову знайти 10 найбільш вживаних слів у тексті та побудувати на їх основі стовпчасту діаграму.

*burgess-busterbrown.txt*

1. **Текст програми**

import nltk

from nltk.corpus import stopwords

import matplotlib.pyplot as plt

from collections import Counter

import string

nltk.download('stopwords')

try:

file = open('burgess-busterbrown.txt', 'r', encoding='utf-8')

except FileNotFoundError:

print("Файл не знайдено!")

exit(0)

def count\_words(text):

sentences = nltk.sent\_tokenize(text)

k\_words = 0

for sentence in sentences:

words = nltk.word\_tokenize(sentence)

k\_words += len(words)

return k\_words

def remove\_stopwords(text):

stop\_words = set(stopwords.words('english'))

cleaned\_text\_arr = [word.lower() for word in text if word.lower() not in stop\_words and word not in string.punctuation]

return ''.join(str(letter) for letter in cleaned\_text\_arr)

def most\_used\_words(text, title):

text1 = text.split()

cnt = Counter(text1)

cort = cnt.most\_common(10)

x = [cort[el][0] for el in range(len(cort))]

y = [cort[el][1] for el in range(len(cort))]

plt.bar(x, y)

plt.title(title)

plt.xlabel("Слова")

plt.ylabel("Кількість у тексті")

plt.show()

text = file.read()

print(f"\nКількість слів у тексті: {count\_words(text)}")

print(f"Кількість слів у тексті (без стоп-слів): {count\_words(text)} \n")

cleaned\_text = remove\_stopwords(text)

most\_used\_words(text, "10 найбільш вживаних слів у тексті")

most\_used\_words(cleaned\_text, "10 найбільш вживаних слів у тексті (без стоп-слів)")

1. **Приклад роботи програми**

**A screenshot of a computer

Description automatically generated**

**A graph of blue bars

Description automatically generated**

**A graph of blue bars

Description automatically generated**

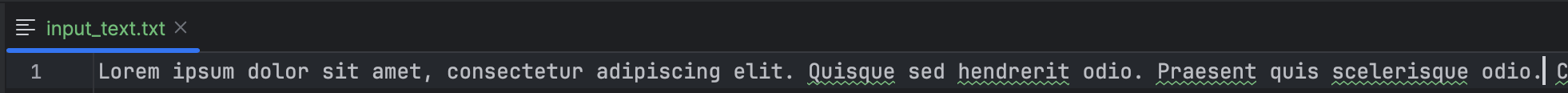
**Завдання 2**

1. **Постановка задачі**

Зберегти у текстовому файлі довільний текст до 100 слів. Виконати наступні дії над текстом:

1. Токенізація по словам
2. Лемматизація та стеммінг
3. Видалення стоп-слів
4. Видалення пункуації

Записати оброблений текст у інший файл. Для виконання завдань лабораторної роботи використати бібліотеку NLTK.



1. **Текст програми**

import nltk

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer, PorterStemmer

import string

nltk.download('punkt')

nltk.download('stopwords')

nltk.download('wordnet')

nltk.download('omw-1.4')

with open('input\_text.txt', 'r') as file:

text = file.read()

tokens = word\_tokenize(text)

lemmatizer = WordNetLemmatizer()

stemmer = PorterStemmer()

lemmas = [lemmatizer.lemmatize(token.lower()) for token in tokens]

stems = [stemmer.stem(token.lower()) for token in tokens]

stop\_words = set(stopwords.words('english'))

filtered\_tokens = [token for token in lemmas if token not in stop\_words]

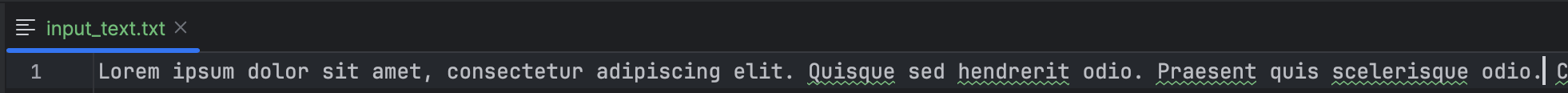
processed\_tokens = [token for token in filtered\_tokens if token not in string.punctuation]

processed\_text = ' '.join(processed\_tokens)

with open('processed\_text.txt', 'w') as file:

file.write(processed\_text)

1. **Приклад роботи програми**

****

****

**Github**

[Посилання](https://github.com/xxehwuq/sumdu/tree/main/2%20%D0%BA%D1%83%D1%80%D1%81/%D0%9F%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%20%D0%BC%D0%BE%D0%B2%D0%BE%D1%8E%20Python/%D0%9B%D0%B0%D0%B1%D0%BE%D1%80%D0%B0%D1%82%D0%BE%D1%80%D0%BD%D0%B0%20%D1%80%D0%BE%D0%B1%D0%BE%D1%82%D0%B0%20%E2%84%9615)