МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

**ЗВІТ З ЛАБОРАТОРНОЇ РОБОТИ №12**

з дисципліни «Програмування мовою Python»

Виконав студент групи ІН-11/2 Штогрін В.О.

Перевірила Парфененко Ю.В.

Варіант 6

Суми 2022

**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.

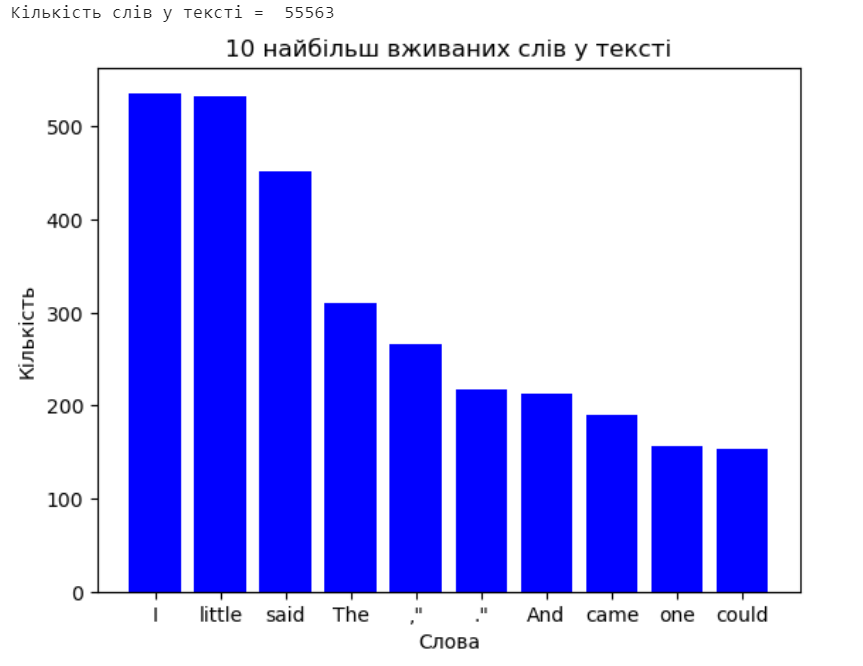
Визначити 10 найбільш вживаних слів у тексті, побудувати на основі цих даних стовпчату діаграму. Виконати видалення з тексту стоп-слів та пунктуації, після чого знову знайти 10 найбільш вживаних слів у тексті та побудувати на їх основі стовпчату діаграму.

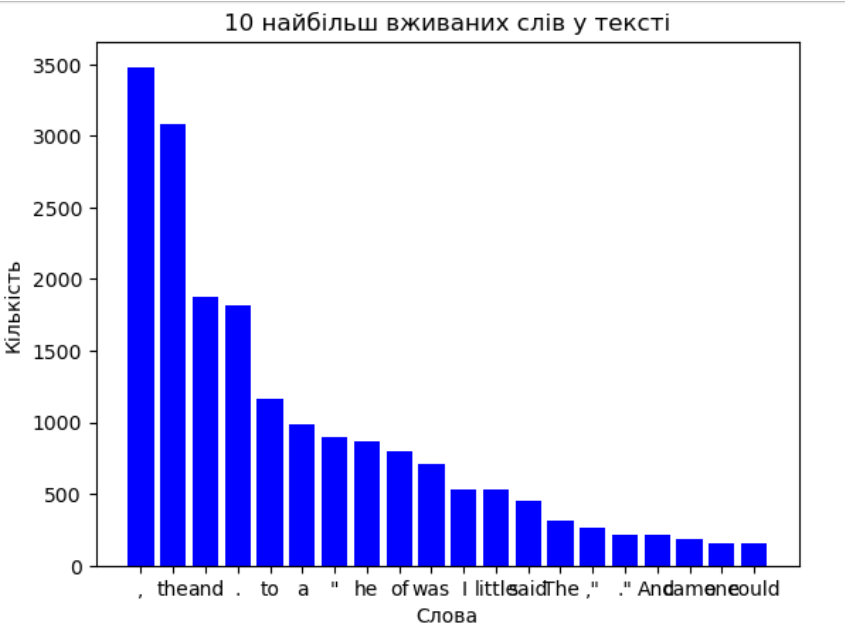


1. **Текст програми**

|  |
| --- |
| import nltk  import matplotlib.pyplot as plt  from collections import Counter  from nltk.corpus import stopwords  nltk.download('gutenberg')  nltk.download('stopwords')  import pandas as pd  import string  def most\_used\_words(bryant):  kol\_words = 0  for sentence in bryant:  kol\_words +=1  kol\_words = Counter(bryant)  cort=kol\_words.most\_common(10)  x = [cort[i][0] for i in range(len(cort))]  y = [cort[i][1] for i in range(len(cort))]  plt.bar(x, y, color = 'blue')  plt.title("10 найбільш вживаних слів у тексті")  plt.xlabel("Слова")  plt.ylabel("Кількість")  plt.show  def mk\_trans\_tab(alice):  return str.maketrans(dict(zip(alice,list(' ' \* len(bryant)))))  def most\_used\_words\_clear\_text(bryant):  stops = stopwords.words('english')  bryant\_new = []  for t in bryant:  if(t not in string.punctuation) and (t not in stops) :  bryant\_new.append(t)  transl\_tab = mk\_trans\_tab(list(string.punctuation) + list ('\r\n«»\-'))  most\_used\_words(bryant\_new)  bryant = nltk.corpus.gutenberg.words('bryant-stories.txt')  print("Кількість слів у тексті = ", len(bryant))  most\_used\_words(bryant)  most\_used\_words\_clear\_text(bryant) |

1. **Приклад роботи програми**





**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  import sys  from nltk.corpus import stopwords  from nltk.stem import PorterStemmer  from nltk.stem import WordNetLemmatizer  import re  from string import punctuation  try:  File = open('nltk.txt', 'r')  except:  print("Файл не знайдено")  sys.exit()  text = File.read()  File.close()  print("Вхідний текст: \n", text, '\n')  stop\_words = set(stopwords.words('english'))  ps = PorterStemmer()  wordnet\_lemmatizer = WordNetLemmatizer()  words = nltk.word\_tokenize(text)  without\_punctuation = [x for x in words if not re.fullmatch('[' + punctuation + ']+', x)]  without\_stop\_words = [word for word in without\_punctuation if not word in stop\_words]  text = ''  for word in without\_stop\_words:  rootWord = ps.stem(word)  lemma = wordnet\_lemmatizer.lemmatize(rootWord)  text1 = ''.join(lemma)  text = text+' '+text1  print("Відредагований текст: \n", text, '\n')  File = open('nltk\_new.txt', 'w')  File.write(text)  print("Файл записано")  File.close() |

1. **Приклад роботи програми**

