

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

ІМЕНІ ІГОРЯ СІКОРСЬКОГО”

Факультет прикладної математики

Кафедра програмного забезпечення комп’ютерних систем

**Лабораторна робота №2**

з дисципліни “Компоненти програмної інженерії”

тема “Збір даних з веб-документів за допомогою мови Python”

|  |  |  |
| --- | --- | --- |
| Виконав  студент I курсу  групи КП-02  Чекурда Андрій Віталійович  (*прізвище, ім’я, по батькові*)  варіант №18 |  |  |

Київ 2020

**Мета роботи**

навчитися отримувати дані з html-сторінок та здійснювати їх

аналіз, використовуючи можливості мови Python

**Постановка завдання**

Реалізуйте програму, яка для довільної сторінки будь-якого сайту новин

буде підраховувати частоту появи слів у тексті новини, частоту появи html-

тегів, кількість посилань та зображень.

#### **Аналіз вимог**

Приклади роботи програми:

|  |
| --- |
| **C:\Users\Admin\PycharmProjects\lab2\venv\Scripts\python.exe C:/Users/Admin/PycharmProjects/lab2/main.py**  **Enter a path to web: www.google.com**  **Enter a word for checking: Google**  **Frequency of word Google in text of news is 0.043478260869565216**  **Enter a tag for checking: a**  **Frequency of tag a is 0.20238095238095238**  **Total number of images is 2**  **Total number of links is 17** |

|  |
| --- |
| **C:\Users\Admin\PycharmProjects\lab2\venv\Scripts\python.exe C:/Users/Admin/PycharmProjects/lab2/main.py**  **Enter a path to web: sadsad**  **url isn't correct** |

**Тексти коду програм**

|  |
| --- |
| main.py |
| import requests  from bs4 import BeautifulSoup  value = input("Enter a path to web: ")  try:  r = requests.get("https://{0}".format(value))  if r.ok:  page = BeautifulSoup(r.text, 'html.parser')  images = page.findAll('img')  links = page.findAll('a')  word = input("Enter a word for checking: ")  numberOfWords = 0  text = page.body.text  text = text.replace('|', '')  text = text.replace('.', '')  text = text.replace('-', '')  text = text.replace('`', '')  text = text.replace('"', '')  text = text.replace('?', '')  text = text.replace('!', '')  text = text.replace(')', '')  text = text.replace('(', '')  text = text.replace(',', '')  newText = text.replace(':', '')  array = newText.split()  for i in range(len(array)):  if array[i] == word:  numberOfWords += 1  frequency1 = numberOfWords / len(array)  print("Frequency of word {0} in text of news is {1}".format(word, frequency1))  tag = input("Enter a tag for checking: ")  numberOfTags = page.findAll('{0}'.format(tag))  numberOfAllTags = page.findAll()  frequency2 = len(numberOfTags) / len(numberOfAllTags)  print("Frequency of tag {0} is {1}".format(tag, frequency2))  print("Total number of images is {0}".format(len(images)))  print("Total number of links is {0}".format(len(links)))  else:  print("url isn't correct")  except:  print("url isn't correct") |

**Висновки**

Виконавши цю лабораторну роботу, я здобув навички користуватись певними інструментами мови python для аналізу веб-сторінок, а саме були використані модуль requests для доступу до даних та обробка даних за допомогою bs4. Також було покращено вміння працювати з рядковим типом даних для обробки тексту, який надійшов з сторінки.