

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-01  Тітов Єгор Павлович  Варіант №17 |  | Перевірив  “\_\_\_\_” “\_\_\_\_\_\_\_\_\_\_\_\_” 20\_\_\_р  викладач  Радченко Костянтин Олександрович |

Київ 2020

**Мета**

Навчитися одержувати дані з html-сторінок та здійснювати їх аналіз, використовуючи можливості мови Python.

**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.

**Код**

|  |
| --- |
| import requests  from bs4 import BeautifulSoup  def find\_word\_frequency(page, seekingWord):  countOfSeekingWords = page.text.count(seekingWord)  return countOfSeekingWords/len(page.text.split()) \* 100  def find\_tag\_frequency(page, seekingTag):  countOfTag = len(page.find\_all(seekingTag))  return countOfTag / len(page.find\_all()) \* 100  def find\_countOf\_links(page):  countOfLinks = 0  for item in page.find\_all():  if item.get('href') == None or item.get('href') == "":  continue  else:  countOfLinks += 1  return countOfLinks  ###########################################################################  inputHtml = input('Enter website link: ')  r = requests.get(inputHtml)  page = BeautifulSoup(r.text, 'html.parser')  ###########################################################################  seekingWord = input('Enter word to seek: ')  wordFrequency = find\_word\_frequency(page, seekingWord)  print(f"- Frequency of word '{seekingWord}' is %.2f %%" % wordFrequency)  ###########################################################################  seekingTag = input('Enter tag to seek: ')  tagFrequency = find\_tag\_frequency(page, seekingTag)  print(f"- Frequency of tag '{seekingTag}' is %.2f %%" % tagFrequency)  ###########################################################################  countOfLinks = find\_countOf\_links(page)  print("- Links: ", countOfLinks)  ###########################################################################  print("- Images: ", len(page.find\_all('img')))  ########################################################################### |

**Хід виконання**

* *За допомогою GET запиту отримаємо вміст шуканої веб-сторінки.*
* *Розпарсимо отриманий запит і дістанемо з нього необхідні дані за допомогою модуля BeautifulSoup4*
* *За допомогою вбудованих в модуль BeautifulSoup4 функцій знайдемо кількість картинок, посилань і частоту появи певного слова і певного тега*

**Результати**

Приклад роботи програми на різних сайтах з різними шуканими значеннями:

|  |
| --- |
| Enter website link: https://tsn.ua/ru  Enter word to seek: Україна  - Frequency of word 'Україна' is 0.00 %  Enter tag to seek: a  - Frequency of tag 'a' is 17.28 %  - Links: 451  - Images: 97 |

|  |
| --- |
| Enter website link: https://www.bbc.com/ukrainian  Enter word to seek: контейнеровоз  - Frequency of word 'контейнеровоз' is 0.29 %  Enter tag to seek: div  - Frequency of tag 'div' is 42.61 %  - Links: 117  - Images: 2 |

**Висновки**

В процесі виконання лабораторної роботи було одержано дані з html-сторінок та здійснено їх аналіз, використовуючи можливості мови Python.

Були використані такі модулі як Requests і BeautifulSoup4, що значно спростило написання коду.