Міністерство освіти і науки України

Центральноукраїнський національний технічний університет

**МЕХАНІКО-ТЕХНОЛОГІЧНИЙ ФАКУЛЬТЕТ**

Кафедра програмування та захисту інформації

**Звіт**

з виконаної лабораторної роботи № 6

дисципліни “ Программування ”

на тему

«Збір даних з веб-документів за допомогою Python»

Виконав :

студент академічної групи КІ-15

Аннаєв А. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Перевірив :

Викладач

Смiрнова Н.В.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Кропивницкий- 2017

**Лабораторная работа №6**

**Тема:** Сбор данных с веб-документов с помощью Python

**Цель:** научиться получать данные с html-страниц и осуществлять их

анализ, используя возможности языка Python

**Задание:**

Реализуйте программу, которая для произвольной страницы любого сайта новостей будет подсчитывать частоту появления слов в тексте новости, частоту появления html-тегов, количество ссылок и изображений

import module

print("\tЛабараторная работа №6")

print("\tПарсинг вебстаниц\n")

URL = 'http://itnews.com.ua/'

result = module.count\_href\_on\_website(URL)

print("Количество ссылок на странице", result)

result = module.count\_image\_on\_website(URL)

print("Количество изображений на странице", result)

result = module.parse\_html(URL)

dictionary = module.counting(result)

print(dictionary)

take\_all\_text = module.take\_text(URL)

for i in take\_all\_text:

print(i)

from bs4 import BeautifulSoup

import urllib.request

import requests

from lxml import html

def count\_href\_on\_website(url):

response = requests.get(url)

parsed\_body = html.fromstring(response.text)

lists = parsed\_body.xpath('//a/@href')

href\_count = len(lists)

return href\_count

def count\_image\_on\_website(url):

response = requests.get(url)

parsed\_body = html.fromstring(response.text)

lists = parsed\_body.xpath('//img/@src')

image\_count = len(lists)

return image\_count

def parse\_html(url):

response = urllib.request.urlopen(url)

html\_text = response.read()

soup = BeautifulSoup(html\_text)

tags\_name = [tag.name for tag in soup.find\_all()]

return tags\_name

def counting(response):

tags\_name = ['meta', 'html', 'body', 'h1', 'h2', 'p',

'span', 'br', 'a', 'img', 'b', 'strong',

'i', 'u', 'ol', 'ul', 'tr', 'td', 'li', 'div',

'h3', 'href']

result = []

for name in tags\_name:

count = response.count(name)

result.append({ name : count})

return result

def take\_text(url):

tags\_name = ['title', 'body', 'h1', 'h2', 'p',

'span', 'br', 'a','ol', 'ul', 'tr', 'td', 'li', 'div',

'h3']

response = urllib.request.urlopen(url).read()

soup = BeautifulSoup(response)

html\_size = len(soup)

result = []

temp = ''

for tag in tags\_name:

for i in soup.find\_all(tag):

temp += i.text

result.append({tag : temp})

return result