Образовательное частное учреждение высшего образования
«Еврейский университет»
Факультет экономики и информатики
Кафедра информатики и математики

ЛАБОРАТОРНАЯ РАБОТА № 4

по дисциплине «программирование на Python»

Выполнил: Высоцкий Р.Н.,
студент 3 курса
03.09.09 Прикладная информатика

Руководитель: доцент кафедры информатики
и математики Демичев Василий Анатольевич,
к. ф.-м. н.

Москва, 2022

Цель: получение практических навыков в использовании регулярных выражений и библиотеке BeautifulSoup.

Код:

```python
import requests
from bs4 import BeautifulSoup
import re

import urllib.parse

def MakeRegFromKeyWords(rawkey):
    rawkey = rawkey.lower()
    key = rawkey.split(" ")

    regexp = r"\s";
    for word in key:
        regexp += word + r"[а-яА-я]*\s+"

    return regexp
```

```python

def ProcessPage(pageurl , depth , regexp , vizited = []):
    lst = []
    if depth <= 0:
        return []
    pageurl = urllib.parse.unquote(pageurl) #удаление служебных элементов из ссылки
    if pageurl.find("http") < 0:
        pageurl = r"http://" + pageurl
    pageurl = pageurl.replace("index.html" , "")
    pageurl = pageurl.rstrip("/")

    if pageurl in vizited:
        return []
    else:
        vizited.append(pageurl)
        print("-->> обрабатываю страницу " , pageurl)
        try:
            resp = requests.get(pageurl)
            soup = BeautifulSoup(resp.text , 'lxml')
            txt = soup.text
            txt = txt.lower()
        except:
            print("--->> ссылка ", pageurl ,"не работает!")
            return []
        match = re.search(regexp , txt)
        if match:
            start = match.start()
            end = match.end()
            print("->> найдено совпадение на " , pageurl)

            s0 = start - 100
            if s0 < 0:
                s0 = 0
            e0 = end + 100
            if e0 >= len(txt):
                e0 = len(txt)-1

            print(txt[s0:e0])
            lst.append([pageurl , start , end , txt[s0:e0]])
```

```
55
56              for tag in soup.find_all("a" , href = True):
57                  mas = ProcessPage(tag['href'] , depth-1 , regexp , vizited)
58                  for items in mas:
59                      lst.append(items)
60
61              return lst
62
63  addres = input("Ведите адрес сайта: ")
64  key = input("Введите ключевую фразу для поиска: ")
65  depth = int(input("Введите поисковую глубину: "))
66  regexp = MakeRegFromKeyWords(key)
67  vizited = []
68  mas = ProcessPage(addres , depth , regexp , vizited)
69  persent = len(mas)/len(vizited) * 100.0
70  print("> процент совпадения ключевой фразы: " , persent)
71  regexp = r"https?\:?\/\/(www\.)?.\w+\.\w+(\.\w+)?(\.\w+)?(\.\w+)?(\.\w+)?"
72  #убираем название сайта
73  websites = {}
74  for item in mas:
75      match = re.search(regexp , item[0])
76      if match:
77          start = match.start()
78          end = match.end()
79          website = item[0][start:end]
80          if website in websites.keys():
81              websites[website] += 1
82          else:
83              websites[website] = 1
84
85  print("наличие информации на сайтах:" , websites)
86
87
```

Программа рекурсивно ищет слова методом из предыдущей лабораторной работы, а так же через BeautifulSoup ищет ссылки на странице и переходит по ним для дальнейшего поиска.