OPEN SOURCE SOFTWARE LAB (15B17CI575)

Lab Assignment 5 (Practice Lab)

Odd 2021

Week 5 & 6: 28 Sept- 9 Oct

Topic Coverage: Python- BeautifulSoup, Panda, MongoDb

Patil Amit Gurusidhappa

19104004

B11

Q1. Extracting URL's from any website

from bs4 import BeautifulSoup

import requests

url = "google.com"

r = requests.get("https://" + url)

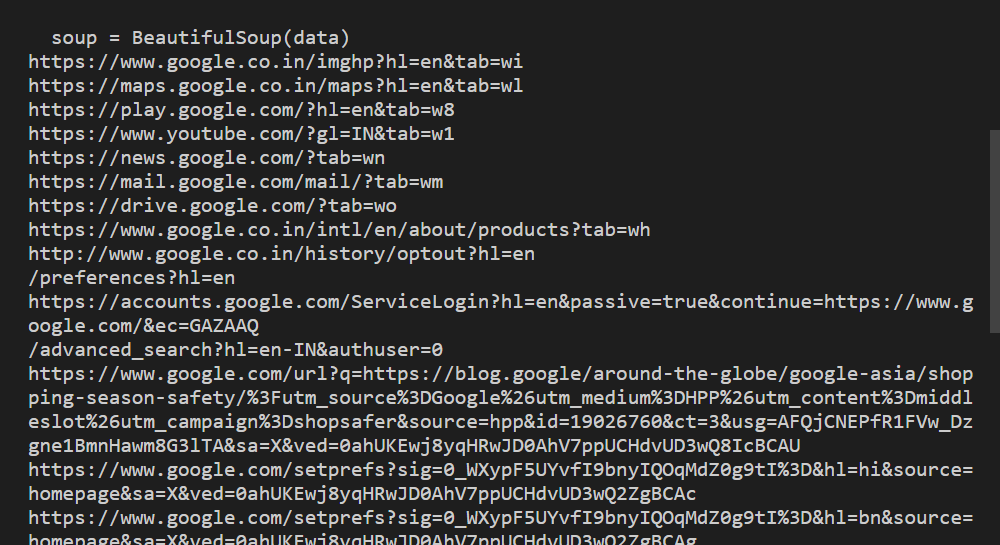
data = r.text

soup = BeautifulSoup(data)

for link in soup.find\_all('a'):

    print(link.get('href'))

Output:



Q2. Extracting Title, Headings, and Links of a website

from bs4 import BeautifulSoup

import requests

req = requests.get(

    'https://en.wikipedia.org/wiki/Python\_(programming\_language)')

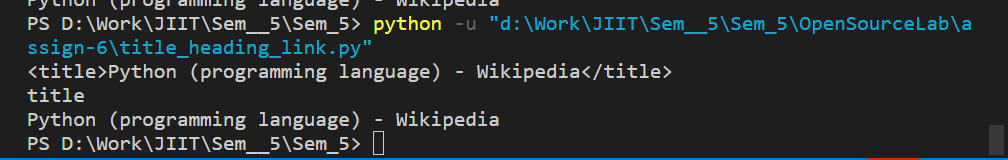
soup = BeautifulSoup(req.text, "lxml")

print(soup.title)

print(soup.title.name)

print(soup.title.string)

Output:



Q3. Extracting the main heading or the first paragraph

from bs4 import BeautifulSoup

import requests

req = requests.get(

    'https://en.wikipedia.org/wiki/Python\_(programming\_language)')

soup = BeautifulSoup(req.text, "lxml")

print(soup.h1)

print(soup.h1.string)

soup.h1['class'] = 'firstHeading, mainHeading'

soup.h1.string.replace\_with("Python - Programming Language")

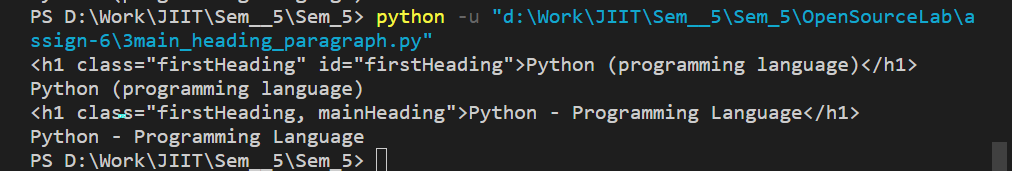
del soup.h1['lang']

del soup.h1['id']

print(soup.h1)

print(soup.h1.string)

Output:



Web Crawling using pyMongo and storing data in MongoDb

from urllib.request import urlopen

from bs4 import BeautifulSoup

from pymongo import Connection

host = 'localhost'

database = 'lotto'

collection = 'mega\_millions'

def mongo\_connection():

    con = Connection(host)

    col = con[database][collection]

    return col

def main():

    col = mongo\_connection()

    page\_num = 1

    total\_pages = 63

    while True:

        if page\_num > total\_pages:

            break

        page\_num = str(page\_num)

        soup = BeautifulSoup(urlopen('http://www.usamega.com/mega-millions-history.asp?p='+page\_num).read())

    for row in soup('table')[4].findAll('tr'):

        win\_dict = {}

        tds = row('td')

    if tds[1].a is not None:

        win\_dict['date'] = tds[1].a.string

    if tds[3].b is not None:

        num\_list = []

    #Told you we would get back to it

    number\_list = tds[3].b.string.split('&middot;')

    for num in number\_list:

        num\_list.append(int(num))

        win\_dict['numbers'] = num\_list

        mega\_number = tds[3].strong.string

        win\_dict['mega\_number'] = int(mega\_number)

        col.insert(win\_dict)

        page\_num = int(page\_num)

        page\_num += 1

main()

Scarping Table and storing using pymongo

from bs4 import BeautifulSoup

import urllib.request

import pymongo

url = 'https://worldpopulationreview.com/country-rankings/olympic-medals-by-country'

page = urllib.request.urlopen(url)

soup = BeautifulSoup(page.read())

soup.prettify()

table = soup.find("div", {"class": "jsx-1911055634"})

cells = []

for row in table.findAll("tr"):

    result = row.findAll("td")

    if len(result) == 6:

        dic = {}

        dic["Country"] = result[0].get\_text()

        dic["Gold"] = result[1].get\_text()

        dic["Silver"] = result[2].get\_text()

        dic["Bronze"] = result[3].get\_text()

        dic["Total"] = result[4].get\_text()

        dic["Population"] = result[5].get\_text()

        cells.append(dic)

print(cells)

myclient = pymongo.MongoClient("mongodb://localhost:27017/")

mydb = myclient["mydatabase"]

mycol = mydb["customers"]

x = mycol.insert\_many(cells)

print(x)

Output

