Date
17/01/22

Neeraj Appari T023

SHETH L.U.J. COLLEGE OF ARTS &
SIR M.V. COLLEGE OF SCIENCE & COMMERCE
Department of Computer Science

Information Retrieval Practical No-5

Aim: Write a program to implement simple Web Crawler.
A) using bs4
B) using scrapy.

Write-ups:

1) Web Crawling - A web crawler is a program or automated script which browses the world wide web in a methological, automated manner. This process is called web crawling or spidering.

2) Bs4 bs4 - Beautiful soup is a python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching and modifying the parse

3) Scrapy - Scrapy is a python framework for large scale web scrapping. It gives all the tools one needs to efficiently extract data from websites, process them as one wants, and store them in one's preferred structure and format

Code

```
import requests
from bs4 import BeautifulSoup
url = "https://www.facebook.com
r = requests.get(url)
htmlContent = r.content
```

Neeraj Appan T073

SHETH L.U.J. COLLEGE OF ARTS &
SIR M.V. COLLEGE OF SCIENCE & COMMERCE
Department of Computer Science

```
soup = BeautifulSoup(htmlContent, 'html.parser')

title = soup.title

paras = soup.find_all('p')

anchors = soup.find_all('a')

anchors = Soup.find_all('a')
all_links = set()
for link in anchas:
    if (link.get('href') != '#'):
        linkText = "https://www.facebook.com" + link.get('href')
        all_links.add(link)
        print(linkText)
```

Information Retreival Pratical-5(A).py - E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-5(A).py (3.9.6)

File Edit Format Run Options Window Help

```
paras = soup.find_all('p')
#print(paras)

#Get all the anchor tags from the page
anchors = soup.find_all('a')
#print(anchors)

#Get first element in the HTML page
#print(soup.find('a'))

#Get classes of any element in the HTML page
#print(soup.find('p')['class'])

#find all the elements with class lead
#print(soup.find_all("p", class_=" "))

#Get the text from the tags/soup
#print(soup.find('p').get_text())
#print(soup.get_text())


#Get all the anchor tags from the page
anchors = soup.find_all('a')
#print(anchors)
all_links = set()
#Get all the links on the page
for link in anchors:
   if(link.get('href') != '#'):
     linkText = "https://www.facebook.com" +link.get('href')
     all_links.add(link)
     print(linkText)
```

Ln: 10 Col: 0

Information Retreival Pratical-5(A).py - E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-5(A).py (3.9.6)

File Edit Format Run Options Window Help

```
import requests
from bs4 import BeautifulSoup
url = "http://facebook.com"
print("Neeraj Appari T073")

#get HTML
r = requests.get(url)
htmlContent=r.content
#print(htmlContent)

#parse HTML
soup= BeautifulSoup(htmlContent,'html.parser')
#print(soup.prettify)

#HTML tree traversal

#Get the title of the HTML Page
title = soup.title
#print(title)

#Get all the paragraphs from the Page
paras  = soup.find_all('p')
#print(paras)

#Get all the anchor tags from the page
anchors = soup.find_all('a')
#print(anchors)

#Get first element in the HTML page
#print(soup.find('a'))

#Get classes of any element in the HTML page
```

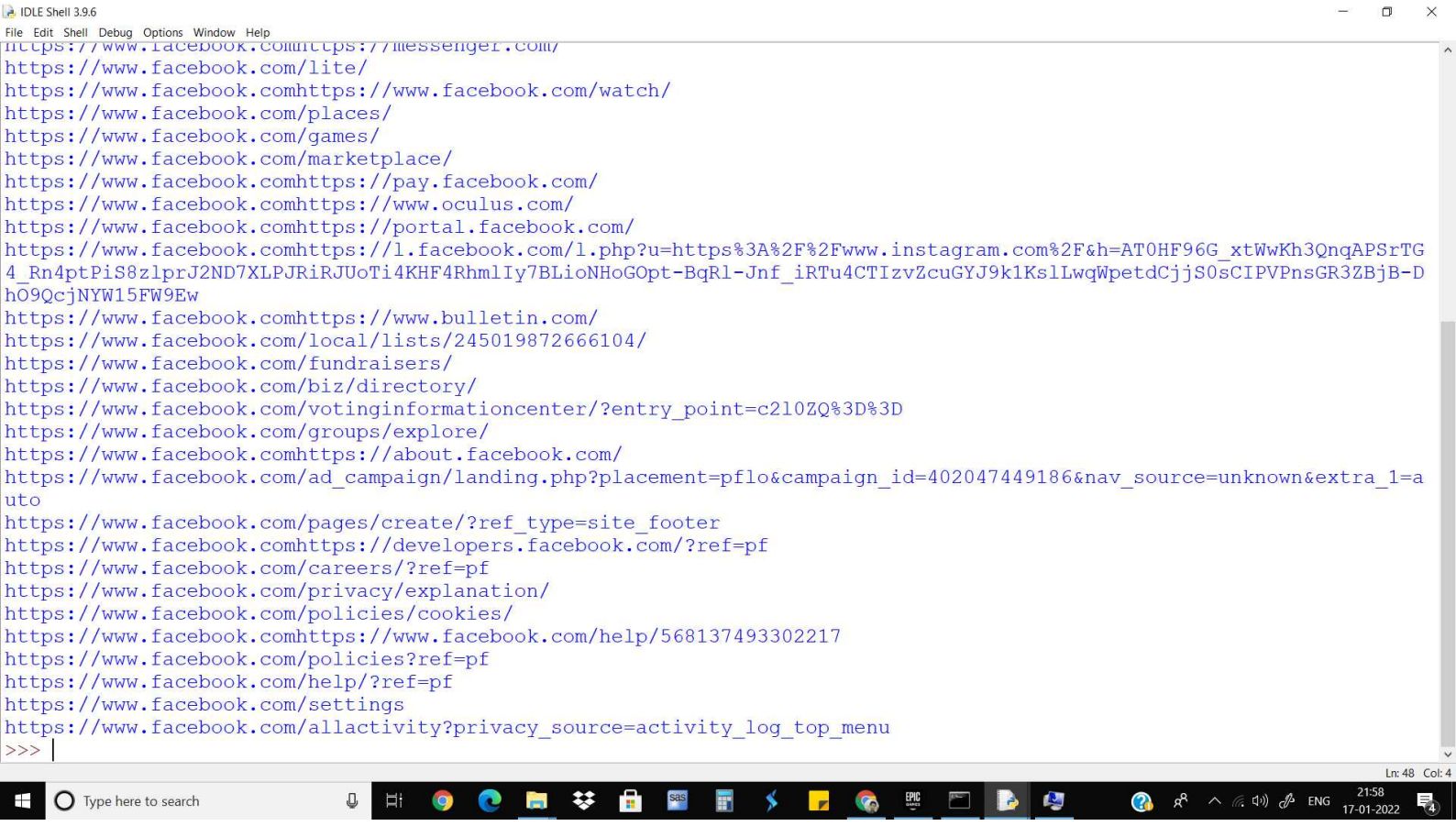
Ln: 10 Col: 0

IDLE Shell 3.9.6

File Edit Shell Debug Options Window Help

```
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-5(A).py
Neeraj Appari T073
https://www.facebook.comhttps://www.facebook.com/recover/initiate/?privacy_mutation_token=eyJ0eXBlIjowLCJjcmVhdGlvbl90aW1lIjoxNjQyNDM2ODg0LCJjYWxsc2l0ZV9pZCI6MzgxMjI5MDc5NTc1OTQ2fQ%3D%3D&ars=facebook_login
https://www.facebook.com/pages/create/?ref_type=registration_form
https://www.facebook.comhttps://hi-in.facebook.com/
https://www.facebook.comhttps://ur-pk.facebook.com/
https://www.facebook.comhttps://gu-in.facebook.com/
https://www.facebook.comhttps://kn-in.facebook.com/
https://www.facebook.comhttps://pa-in.facebook.com/
https://www.facebook.comhttps://ta-in.facebook.com/
https://www.facebook.comhttps://bn-in.facebook.com/
https://www.facebook.comhttps://te-in.facebook.com/
https://www.facebook.comhttps://ml-in.facebook.com/
https://www.facebook.comhttps://en-gb.facebook.com/
https://www.facebook.com/reg/
https://www.facebook.com/login/
https://www.facebook.comhttps://messenger.com/
https://www.facebook.com/lite/
https://www.facebook.comhttps://www.facebook.com/watch/
https://www.facebook.com/places/
https://www.facebook.com/games/
https://www.facebook.com/marketplace/
https://www.facebook.comhttps://pay.facebook.com/
https://www.facebook.comhttps://www.oculus.com/
https://www.facebook.comhttps://portal.facebook.com/
https://www.facebook.comhttps://l.facebook.com/l.php?u=https%3A%2F%2Fwww.instagram.com%2F&h=AT0HF96G_xtWwKh3QnqAPSrTG4_Rn4ptPiS8zlprJ2ND7XLPJRiRJUoTi4KHF4RhmlIy7BLioNHoGOpt-BqRl-Jnf_iRTu4CTIzvZcuGYJ9k1KslLwqWpetdCjjS0sCIPVPnsGR3ZBjB-DbQ9QciNYW15EW9Ew
```

Ln: 48 Col: 4


IDLE Shell 3.9.6
File Edit Shell Debug Options Window Help
https://www.facebook.comhttps://messenger.com/
https://www.facebook.com/lite/
https://www.facebook.comhttps://www.facebook.com/watch/
https://www.facebook.com/places/
https://www.facebook.com/games/
https://www.facebook.com/marketplace/
https://www.facebook.comhttps://pay.facebook.com/
https://www.facebook.comhttps://www.oculus.com/
https://www.facebook.comhttps://portal.facebook.com/
https://www.facebook.comhttps://l.facebook.com/l.php?u=https%3A%2F%2Fwww.instagram.com%2F&h=AT0HF96G_xtWwKh3QnqAPSrTG4_Rn4ptPiS8zlprJ2ND7XLPJRiRJUoTi4KHF4RhmlIy7BLioNHoGOpt-BqRl-Jnf_iRTu4CTIzvZcuGYJ9k1KslLwqWpetdCjjS0sCIPVPnsGR3ZBjB-DhO9QcjNYW15FW9Ew
https://www.facebook.comhttps://www.bulletin.com/
https://www.facebook.com/local/lists/245019872666104/
https://www.facebook.com/fundraisers/
https://www.facebook.com/biz/directory/
https://www.facebook.com/votinginformationcenter/?entry_point=c2l0ZQ%3D%3D
https://www.facebook.com/groups/explore/
https://www.facebook.comhttps://about.facebook.com/
https://www.facebook.com/ad_campaign/landing.php?placement=pflo&campaign_id=402047449186&nav_source=unknown&extra_1=auto
https://www.facebook.com/pages/create/?ref_type=site_footer
https://www.facebook.comhttps://developers.facebook.com/?ref=pf
https://www.facebook.com/careers/?ref=pf
https://www.facebook.com/privacy/explanation/
https://www.facebook.com/policies/cookies/
https://www.facebook.comhttps://www.facebook.com/help/568137493302217
https://www.facebook.com/policies?ref=pf
https://www.facebook.com/help/?ref=pf
https://www.facebook.com/settings
https://www.facebook.com/allactivity?privacy_source=activity_log_top_menu
>>>
Ln: 48 Col: 4
Type here to search
ENG
21:58
17-01-2022