Module 8: Computer Vision OpenCV and Visualisation using Bokeh

Case Study Solution

Case Study Solution

1. Write a program to fetch hyperlinks from any website which user enters.

Solution

```
from bs4 import BeautifulSoup
import requests
url = input("Enter a website to extract the URL's from: ")
r  = requests.get("http://" +url)
data = r.text
soup = BeautifulSoup(data)
for link in soup.find_all('a'):
   print(link.get('href'))
```

2. Write a program to download all the videos from youtube.com for django from the hyperlink given below

https://www.youtube.com/playlist?list=PLxxA5z-8B2xk4szCgFmgonNcCboyNneMD

Solution

```
import youtube_dl
from bs4 import BeautifulSoup as BS, SoupStrainer as SS
import requests

course_code = 'PLxxA5z-8B2xk4szCgFmgonNcCboyNneMD'#'PL385A53B00B8B158E'
url = 'https://www.youtube.com/playlist?list='+course_code
resp = requests.get(url)
print(resp.url)
html = resp.content
resp.close()

ss = SS('tr')
soup = BS(html, 'html.parser', parse_only=ss)

base_url = 'https://www.youtube.com/watch?v={0}&index={1}&list='+course_code
fn1 = lambda tag:tag.has_attr('data-video-id') and tag.has_attr('class') and "yt-uix-
tile" in tag.attrs['class']
cnt = 1
for tag in soup.find_all(fn1):
   #print tag.attrs#.tr.attrs
```

```
url = base_url.format(str(tag.attrs['data-video-id']),str(cnt))
print(url)
cnt += 1
youtube_dl.YoutubeDL().download([url])
```

3. Create a csv file with name and hyperlink after fetching it from the web page

http://bioguide.congress.gov/biosearch/biosearch.asp

Select any of the option from it as in the below screenshot and click on search

Later download the page source, save it in html file and then perform scrapping.

Solution

```
from bs4 import BeautifulSoup
import csv
soup = BeautifulSoup (open("test.html"),'lxml')
f = csv.writer(open("outfile.csv", "w"))
f.writerow(["Name", "Link"]) # Write column headers as the first line
links = soup.find_all('a')
for link in links:
 names = link.contents[0]
 fullLink = link.get('href')
 f.writerow([names,fullLink])
```

```
f = open("outfile.csv", "r")
f.read()
```

4. from the question above, fetch only the hyperlinks

Solution

```
from bs4 import BeautifulSoup
import csv
soup = BeautifulSoup (open("test.html"),'lxml')
trs = soup.find_all('tr')
for tr in trs:
 for link in tr.find_all('a'):
    fulllink = link.get ('href')
    print(fulllink) #print in terminal to verify results
```

5. Write Perform the web scrapping on the following page

```
<html>
 <head>
  <title>
   Page title
  </title>
 </head>
 <body>
  <p id="firstpara" align="center">
   This is paragraph
   <b>
    one
   </b>
  </p>
  <p id="secondpara" align="blah">
   This is paragraph
   <b>
    two
   </b>
  </p>
 </body>
</html>
```

i) Read the page using BeautifulSoup and show it in well formatted indented manner.

ii) Print the b tag from the page
iii) Print all the tags that starts from b
iv) Print text from the tags having 'title' and 'p'. by using lists
v) Print text from the tags having 'title' and 'p'. by using dictionaries
vi) Print all the tag names present in the page
vii) Print the complete tag that have two, and only two, attributes
viii) Print the tags that have one-character names and no attributes
ix) Print all the tags which have a value of "center" for them "align" attribute
x) From the xml content
'<person name="Bob"><parent rel="mother" name="Alice">'
Print the attributes having "name" as "Alice"

Solution

```
#Ans i)
from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
    '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
    '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
    '</html>']
soup = BeautifulSoup("".join(doc), 'lxml')
print(soup.prettify())

#Ans ii)

from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
    '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
    '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
    '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
print(soup.findAll('b'))

#Ans iii)

from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
    '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
    '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
    '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
import re
tagsStartingWithB = soup.findAll(re.compile('^b'))
```

```
for tag in tagsStartingWithB:
   print(tag.name )

#Ans iv)
from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
      '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
      '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
      '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
found_tags= soup.findAll(['title', 'p'])
for tag in found_tags:
   print(tag.text)

#Ans v)

from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
      '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
      '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
      '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
found_tags= soup.findAll({'title' : True, 'p' : True})
for tag in found_tags:
   print(tag.text)

#Ans vi)

from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
      '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
      '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
      '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
allTags = soup.findAll(True)
for tag in allTags:
   print(tag.name)

#Ans vii)
from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
      '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
      '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
      '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
```

```
allTags = soup.findAll(lambda tag: len(tag.attrs) == 2)
for tag in allTags:
   print(tag)
```

#Ans viii)

```
from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
      '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
      '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
      '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
allTags=soup.findAll(lambda tag: len(tag.name) == 1 and not tag.attrs)
for tag in allTags:
   print(tag)
```

#Ans ix)

```
from bs4 import BeautifulSoup
doc = ['<html><head><title>Page title</title></head>',
      '<body><p id="firstpara" align="center">This is paragraph <b>one</b>.',
      '<p id="secondpara" align="blah">This is paragraph <b>two</b>.',
      '</html>']
soup = BeautifulSoup(''.join(doc), 'lxml')
allTags=soup.findAll(align="center")
for tag in allTags:
   print(tag)
```

#Ans x)

```
from bs4 import BeautifulSoup
xml = '<person name="Bob"><parent rel="mother" name="Alice">'
xmlSoup = BeautifulSoup(''.join(xml), 'lxml')
xmlSoup.findAll(name="Alice")
found=xmlSoup.findAll(attrs={"name" : "Alice"})
for tag in found:
   print(tag)
```