CSL558: Machine Learning

Instructor : [Dr. Chandra Prakash]

- For more information visit the class website.

LAB Assignment 3: Data Collection Using Web Scraping

Assigning Date : 18-01-2021

Due Date: 24-Jan-2021

Student Name: Rohit Byas

Student Roll No.: 181210043

Assignment Instructions

You must save your as Assignment_NO_Yourname

Your source file will most likely end in.pynbif you are using a Jupyter notebook; however, itmight also end in.pyif you are using a Python script.

You have to add your name and roll no in the Google Colab Instructions section below and printit.

Google CoLab Instructions

The following code ensures that Google CoLab is running the correct version of TensorFlow.

```
try:
  from google.colab import drive
  %tensorflow_version 2.x
  COLAB = True
  print("Hello World")
  print("Note: using Google CoLab")
except:
  print("Hello NITD")
  print("Note: not using Google CoLab")
  COLAB = False

# Print your name and Roll No.
print('Rohit Byas')
print('181210043')
```

```
Hello World
Note: using Google CoLab
Rohit Byas
181210043
```

Task 1:

1. Collect the first 100 quotes from the website http://quotes.toscrape.com/
2. The output should store in three column namely- Quote, Author and Tags.

   Hint:

- You may use panda Dataframe framework as pd.DataFrame for storing the data
- use urllib package

```
!pip install BeautifulSoup4
```

```
Requirement already satisfied: BeautifulSoup4 in /usr/local/lib/python3.6/dis
```

```
!pip install tqdm
```

```
Requirement already satisfied: tqdm in /usr/local/lib/python3.6/dist-packages
```

```
# import different libraries
import json # for json data
import pandas as pd # for data analysis and manipulation
from bs4 import BeautifulSoup # for parsing HTML
from urllib.request import urlopen, Request # for http requests
from IPython.display import display
from ipywidgets import Checkbox


# We need to add agree to legal and ethical concerns before scrapping
box = Checkbox(False, indent=False)
display(box, f"I, {input('Enter your name: ')}, agree to the above Legal and Ethic
```

```
Enter your name: Rohit
```

```
'I, Rohit, agree to the above Legal and Ethical concerns. If I do anything,u
nethical I will be responsible for it.'
```

```
# Defining header that will be send while doing http requests
hdr = {
  'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTI
  'From': 'nitdelhi.ac.in',
  'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
  'Accept-Charset': 'ISO-8859-1,utf-8;q=0.7,*;q=0.3',
  'Accept-Encoding': 'none',
  'Accept-Language': 'en-US,en;q=0.8',
}
```

```
}


# On each page there are 10 quotes, so we need to scrap 10 pages
data = []
for i in range(1,11):

  # get request to the website with page number and header as the link of website
  request=Request(f"http://quotes.toscrape.com/page/{i}/", headers=hdr)

  # reading and decoding the response of request
  html=urlopen(request).read().decode()

  # parsing the webpage by using beautiful soup
  soup=BeautifulSoup(html,'html.parser')

  # quotes are present in div having class "quote"
  quotes = soup.find_all('div', class_='quote')

  #iterate throug those divs
  for quote in quotes:
      # finding the quote with span tag and text class
      text = quote.find('span', class_='text').text
      # finding the author with small tag and author class
      author = quote.find('small', class_='author').text
      # initialising an empty list for tags
      tags = []
      # finding all the a tags with tag class
      tag = quote.find_all('a', class_='tag')
      #iterating through the tags
      for t in tag:
        tags.append(t.text)
      # appending the quotes found on this page to scraped list
      data.append([text, author,tags])

# Creatind dataframe consisting of scrapped qoutes, authorm tags
DataFrame = pd.DataFrame(data,columns=['Quote','Author','Tags'])
DataFrame
```

	Quote	Author	Tags
0	“The world as we have created it is a process ...	Albert Einstein	[change, deep-thoughts, thinking, world]
	“It is our choices, Harry, that show what		

Supplementary Problem :

1. Add two more column of Date of Birth (DoB) and Place of Birth(PoB) of the Author to the output

4	Imperfection is beauty, madness is	Marilyn	[be-yourself, inspirational]

```
# On each page there are 10 quotes, so we need to scrap 10 pages
data = []
for i in range(1,11):

  # get request to the website with page number and header as the link of website
  request=Request(f"http://quotes.toscrape.com/page/{i}/", headers=hdr)

  # reading and decoding the response of request
  html=urlopen(request).read().decode()

  # parsing the webpage by using beautiful soup
  soup=BeautifulSoup(html,'html.parser')

  # quotes are present in div having class "quote"
  quotes = soup.find_all('div', class_='quote')

  #iterate throug those divs
  for quote in quotes:
      # finding the quote with span tag and text class
      text = quote.find('span', class_='text').text
      # finding the author with small tag and author class
      author = quote.find('small', class_='author').text

      # find the link to about author page
      link = quote.find('a',href=True).attrs['href']

      # a get request to about author page
      request=Request(f"http://quotes.toscrape.com{link}/", headers=hdr)
      html=urlopen(request).read().decode()

      # parsing the webpage by using beautiful soup
      soup=BeautifulSoup(html,'html.parser')

      # finding the author dob and pob
      dob = soup.find('span', class_='author-born-date').text
      pob = soup.find('span', class_='author-born-location').text
      pob = pob.replace('in ','')

      # initialising an empty list for tags
      tags = []
      # finding all the a tags with tag class
```

```
        tag = quote.find_all('a', class_='tag')
        #iterating through the tags
        for t in tag:
          tags.append(t.text)
        # appending the quotes found on this page to scraped list
        data.append([text, author,tags, dob, pob])

    # Creatind dataframe consisting of scrapped qoutes, authorm tags
    DataFrame = pd.DataFrame(data,columns=['Quote','Author','Tags', 'Date of Birth(Autl
    DataFrame
```

	Quote	Author	Tags	Date of Birth(Author)	Place of Birth (Author)
0	“The world as we have created it is a process ...	Albert Einstein	[change, deep-thoughts, thinking, world]	March 14, 1879	Ulm, Germany
1	“It is our choices, Harry, that show what we t...	J.K. Rowling	[abilities, choices]	July 31, 1965	Yate, South Gloucestershire, England, The Unit...
2	“There are only two ways to live your life. On...	Albert Einstein	[inspirational, life, live, miracle, miracles]	March 14, 1879	Ulm, Germany
3	“The person, be it gentleman or lady, who has ...	Jane Austen	[aliteracy, books, classic, humor]	December 16, 1775	Steventon Rectory, Hampshire, The United Kingdom
4	“Imperfection is beauty, madness is genius and...	Marilyn Monroe	[be-yourself, inspirational]	June 01, 1926	The United States
...	...	...	...	...	...
95	“You never really understand a person until yo...	Harper Lee	[better-life-empathy]	April 28, 1926	Monroeville, Alabama, The United States
96	“You have to write the book that wants to be w...	Madeleine L'Engle	[books, children, difficult, grown-ups, write,...	November 29, 1918	New York City, New York, The United States