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: Raghav Shukla

Student Roll No.: 181210038

Agenda for the Assignment 3

1. Understand the legal and Ethical issue with web-scrapping. In most of the cases scrapping is legal until you intentionally crash the website or use the data for commercial purposes. You should check the size of the webiste if you are going to scrape the entire website, use google [1]. You can scrape any kind of data that is publicly available but if the data is copyrighted then you cannot use it as your own [2].
2. Collect the data from the website http://quotes.toscrape.com/

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

# Printing the name and Roll No.
print('Raghav Shukla')
print('181210038')

# Printing the curent time
import datetime
print(datetime.datetime.now())
```

```
Hello World
Note: using Google CoLab
Raghav Shukla
181210038
2021-01-24 17:10:39.952258
```

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/dist-packages (4.6.3)
```

```
!pip install tqdm
```

```
Requirement already satisfied: tqdm in /usr/local/lib/python3.6/dist-packages (4.41.1)
```

```
# importing different types of libraries
import json # for json data
import pandas as pd # for data analysis & 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 must 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 Ethical concerns. If I do anything,unethical I will be r
```

```
Enter your name: Raghav Shukla
```

☐

```
'I, Raghav Shukla, agree to the above Legal and Ethical concerns. If I do anything,unethical 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 (KHTML, like Gecko) Chrome/87.0.4280.66 Safari/537.36',
  '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',
}
```

```
# We need to scrap 10 pages as there are 10 quotes.
```

```
data = []
for i in range(1,11):

  # get request to the website with page number and header as the link of website ends with /page{page_no}
  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 & text class
      text = quote.find('span', class_='text').text
      # finding the author which has small tag & author class
      author = quote.find('small', class_='author').text
      # initialising an empty list for the tags
      tags = []
      # finding all tags 'a' with tag class
      tag = quote.find_all('a', class_='tag')
      #iterating through the tags
      for t in tag:
        tags.append(t.text)
      # appending those quotes which were found on this page to the scraped list
      data.append([text, author,tags])

# Create a dataframe which consists of scrapped qoutes and 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]
1	“It is our choices, Harry, that show what we t...	J.K. Rowling	[abilities, choices]
2	“There are only two ways to live your life. On...	Albert Einstein	[inspirational, life, live, miracle, miracles]
3	“The person, be it gentleman or lady, who has ...	Jane Austen	[aliteracy, books, classic, humor]
4	“Imperfection is beauty, madness is genius and...	Marilyn Monroe	[be-yourself, inspirational]
...	...	...	...
95	“You never really understand a person until yo...	Harper Lee	[better-life-empathy]
96	“You have to write the book that wants to be w...	Madeleine L'Engle	[books, children, difficult, grown-ups, write,...
97	“Never tell the truth to people who are not wo...	Mark Twain	[truth]
98	“A person's a person, no matter how small.”	Dr. Seuss	[inspirational]

Supplementary Problem :

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

```
# We need to scrap 10 pages as there are 10 quotes.
data = []
for i in range(1,11):

  # get request to the website with page number and header as the link of website ends with /page{page_no}
  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

        # finding 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()

        # Using the beautiful app to parse the webpage
        soup=BeautifulSoup(html,'html.parser')

        # finding the dob & pob of author
        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 tags 'a' 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(Author)','Place of Birth (Author)'])
```

```
DataFrame = pd.DataFrame(data,columns=[ 'Quote' , 'Author' , 'Tags' , 'Date of Birth(Author)' , 'Place of Birth (Author)' ])
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