Lab Assignment (Practice Lab)
Topic Coverage: Beautiful Soup, Pandas and MongoDB (Part 1)
Week: 13th Oct – 20th Oct

Beautiful Soup

Beautiful Soup is a **Python package for parsing HTML and XML documents** (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites.

Follow the link given below for basic understanding of web scraping and how can we do it with beautiful soup in Python:

https://www.dataquest.io/blog/web-scraping-python-using-beautiful-soup/

Extracting data using beautiful soup (refer to anyone of the following):

https://analyticsindiamag.com/beautiful-soup-webscraping-python/

https://programminghistorian.org/en/lessons/retired/intro-to-beautiful-soup

https://www.analyticsvidhya.com/blog/2021/08/a-simple-introduction-to-web-scraping-with-beautiful-soup/

For more in-depth understanding refer to:

https://www.tutorialspoint.com/beautiful_soup/index.htm
https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/

Quick Summary:

- How to install beautifulsoup library and pip using requests in python
  - $ pip install requests
  - $ pip install beautifulsoup4
- How to install beautifulsoup library using setup.py in python
  - $ python setup.py install
- How to create a soup using HTML parser
  - from bs4 import BeautifulSoup
    soup = BeautifulSoup("<html> This is <b>invalid HTML </html>", "html.parser")
  - Extracting URL's from any website

- ```
  from bs4 import BeautifulSoup
  import requests
  url = raw_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'))
  ```

- The BeautifulSoup object can accept two arguments. The first argument is the actual markup, and the second argument is the parser that you want to use. The different parsers are: html.parser, lxml, and html5lib. The lxml parser has two versions, an HTML parser and an XMLparser.

- Extracting Title, Headings, and Links of a website

- ```
  import requests
  from bs4 import BeautifulSoup
  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)
  ```

- Extracting the main heading or the first paragraph

- ```
  print(soup.h1)
  print(soup.h1.strin)
  ```

- ```
  soup.h1['class'] = 'firstHeading, mainHeading'
  soup.h1.string.replace_with("Python - Programming Language")
  del soup.h1['lang']
  del soup.h1['id']
  ```

Practice Questions on Beautiful Soup in Python

1. Scrap the data from the following URL's

```
'http://www.reuters.com/finance/stocks/company-officers/GOOG.O',
'http://www.reuters.com/finance/stocks/company-officers/AMZN',
'http://www.reuters.com/finance/stocks/company-officers/AAPL'
```

2. Loop through these URLs, scrape table, pass information to array to print the following data.

```
'Name', 'Age', 'Year_Joined', 'Title/Position'
```

Sample output:

	URL	Name	Age	Year_Joined	Title
0	http://www.reuters.com/finance/stocks/company-...	Eric Schmidt	61	2015	Executive Chairman of the Board of Director
1	http://www.reuters.com/finance/stocks/company-...	Sergey Brin	43	2015	President, Director
2	http://www.reuters.com/finance/stocks/company-...	Lawrence Page	44	2015	Chief Executive Officer, Director

3. Create new array, check length to ensure things pulled in correctly.

4. Finally export data to CSV.

Submit csv file also along with other snapshots.

Pandas

Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.

Follow the following links:

https://www.w3schools.com/python/pandas/default.asp

https://www.geeksforgeeks.org/pandas-tutorial/

The following tutorials will provide you with step-by-step instructions on how to work with Pandas:

https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/

Practice Questions on Pandas in Python

Use the 'Automobile_data.csv' and answer following questions by using Pandas library:

1. From given data set, print first and last five rows.

2. Replace all column values which contain ‘?’ and ‘n.a.’ with NaN. Update the CSV file.

3. Print all BMW car details.

4. Count total cars per company.

5. Find each company’s Highest price car.

6. Find the average mileage of each car making company.

7. Merge two data frames using the following condition:
   Create two data frames using the following two Dicts, Merge two data frames, and append second data frame as a new column to the first data frame.
   - Car_Price = {'Company': ['Toyota', 'Honda', 'BMW', ' Audi '], 'Price': [23845, 17995, 135925, 71400]}
   - Car_Horsepower = {'Company': ['Toyota', 'Honda', 'BMV', 'Audi'], 'horsepower': [141, 80, 182, 160]