labassignment5

February 19, 2023

1 Lab Assignment 5: Web Scraping

1.1 DS 6001: Practice and Application of Data Science

1.1.1 Instructions

Please answer the following questions as completely as possible using text, code, and the results of code as needed. Format your answers in a Jupyter notebook. To receive full credit, make sure you address every part of the problem, and make sure your document is formatted in a clean and professional way.

For the following problems, you will be scraping http://books.toscrape.com/. This website is a fake book retailer, designed to mimic the design of many retail websites. It exists solely to help students practice web-scraping, so there aren't going to be any ethical concerns with this particular exercise, and there shouldn't be any issues with rate limits or other gates that could prevent web-scraping. Take a moment and look at this website, so that you know what you will be working with.

Your goal is to generate a dataframe with four columns: one for the title, one for the price, one for the star-rating, and one for the book cover JPEG's URL. The dataframe will have 1000 rows, one for each of the 1000 books listed on the 50 pages of this website.

1.2 Problem 0

Import the following libraries:

[1]:
```python
import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup
import sys
sys.tracebacklimit = 0 # turn off the error tracebacks
```

1.3 Problem 1

Pull the HTML code from http://books.toscrape.com/. Make sure you provide a user agent string. Then parse this HTML code and save the parsed code as a separate Python variable. [3 points]

[2]:
```python
headers = {
    'user-agent': 'Lab Assignment 5: Web Scraping, version 0.1.0␣
 ↪(tsl2b@virginia.edu) (Language=Python 3.9.13; Platform=Mac OSX 13.1)'}
```

```
root = 'http://books.toscrape.com'
response = requests.get(root, headers = headers)
text = response.text
beautifulSoup = BeautifulSoup(text, 'html')
```

1.3.1 Problem 2

Extract all 20 of the book titles and save them in a list. [2 points]

[3]:
```
list_of_anchors = beautifulSoup.find_all('a', title = True)
list_of_titles = [a['title'] for a in list_of_anchors]
print(len(list_of_titles))
list_of_titles
```

```
20
```

[3]:
```
['A Light in the Attic',
 'Tipping the Velvet',
 'Soumission',
 'Sharp Objects',
 'Sapiens: A Brief History of Humankind',
 'The Requiem Red',
 'The Dirty Little Secrets of Getting Your Dream Job',
 'The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria
Woodhull',
 'The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936
Berlin Olympics',
 'The Black Maria',
 'Starving Hearts (Triangular Trade Trilogy, #1)',
 "Shakespeare's Sonnets",
 'Set Me Free',
 "Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)",
 'Rip it Up and Start Again',
 'Our Band Could Be Your Life: Scenes from the American Indie Underground,
1981-1991',
 'Olio',
 'Mesaerion: The Best Science Fiction Stories 1800-1849',
 'Libertarianism for Beginners',
 "It's Only the Himalayas"]
```

1.3.2 Problem 3

Extract the price of each of the 20 books and save these prices in a list. (The prices are listed in British pounds, and include the £ symbol. Remove the £ symbols: if you've saved the prices in a list named `prices`, then the following code should work: `prices = [s.replace('Â£', '') for s in prices]`.) [2 points]

[4]:

```
list_of_paragraphs = beautifulSoup.find_all('p', 'price_color')
list_of_prices = [p.string.replace('Â£', '') for p in list_of_paragraphs]
print(len(list_of_prices))
list_of_prices
```

```
20
```

[4]:

```
['51.77',
 '53.74',
 '50.10',
 '47.82',
 '54.23',
 '22.65',
 '33.34',
 '17.93',
 '22.60',
 '52.15',
 '13.99',
 '20.66',
 '17.46',
 '52.29',
 '35.02',
 '57.25',
 '23.88',
 '37.59',
 '51.33',
 '45.17']
```

1.4 Problem 4

Extract the star level ratings for the 20 books. [Hint: for tags such as `<p class="star-rating One">` in which the class has a space, the class is actually a list in which the first item in the list is `"star-rating"` and the second item in the list is `"One"`. It's possible to search on either item in this list.] [3 points]

[5]:

```
list_of_paragraphs = beautifulSoup.find_all('p', 'star-rating')
list_of_ratings = [p['class'][1] for p in list_of_paragraphs]
print(len(list_of_ratings))
list_of_ratings
```

```
20
```

[5]:

```
['Three',
 'One',
 'One',
 'Four',
 'Five',
 'One',
 'Four',
 'Three',
 'Four',
 'One',
 'Two',
 'Four',
 'Five',
 'Five',
 'Five',
 'Three',
 'One',
 'One',
 'Two',
 'Two']
```

1.5 Problem 5

Extract the URLs for the JPEG thumbnail images that show the covers of the 20 books. (Maybe we want to mine the images to build models that predict the star level, literally judging books by their covers.) [2 points]

[6]:

```
list_of_images = beautifulSoup.find_all('img')
list_of_urls = [root + img['src'] for img in list_of_images]
list_of_urls
```

[6]:

```
['http://books.toscrape.commedia/cache/2c/da/2cdad67c44b002e7ead0cc35693c0e8b.jp
g',
 'http://books.toscrape.commedia/cache/26/0c/260c6ae16bce31c8f8c95daddd9f4a1c.jp
g',
 'http://books.toscrape.commedia/cache/3e/ef/3eef99c9d9adef34639f510662022830.jp
g',
 'http://books.toscrape.commedia/cache/32/51/3251cf3a3412f53f339e42cac2134093.jp
g',
 'http://books.toscrape.commedia/cache/be/a5/bea5697f2534a2f86a3ef27b5a8c12a6.jp
g',
 'http://books.toscrape.commedia/cache/68/33/68339b4c9bc034267e1da611ab3b34f8.jp
g',
 'http://books.toscrape.commedia/cache/92/27/92274a95b7c251fea59a2b8a78275ab4.jp
g',
 'http://books.toscrape.commedia/cache/3d/54/3d54940e57e662c4dd1f3ff00c78cc64.jp
g',
 'http://books.toscrape.commedia/cache/66/88/66883b91f6804b2323c8369331cb7dd1.jp
g',
 'http://books.toscrape.commedia/cache/58/46/5846057e28022268153beff6d352b06c.jp
g',
 'http://books.toscrape.commedia/cache/be/f4/bef44da28c98f905a3ebec0b87be8530.jp
g',
 'http://books.toscrape.commedia/cache/10/48/1048f63d3b5061cd2f424d20b3f9b666.jp
g',
 'http://books.toscrape.commedia/cache/5b/88/5b88c52633f53cacf162c15f4f823153.jp
g',
 'http://books.toscrape.commedia/cache/94/b1/94b1b8b244bce9677c2f29ccc890d4d2.jp
g',
 'http://books.toscrape.commedia/cache/81/c4/81c4a973364e17d01f217e1188253d5e.jp
g',
 'http://books.toscrape.commedia/cache/54/60/54607fe8945897cdcced0044103b10b6.jp
g',
 'http://books.toscrape.commedia/cache/55/33/553310a7162dfbc2c6d19a84da0df9e1.jp
g',
 'http://books.toscrape.commedia/cache/09/a3/09a3aef48557576e1a85ba7efea8ecb7.jp
g',
 'http://books.toscrape.commedia/cache/0b/bc/0bbcd0a6f4bcd81ccb1049a52736406e.jp
g',
 'http://books.toscrape.commedia/cache/27/a5/27a53d0bb95bdd88288eaf66c9230d7e.jp
g']
```

1.6 Problem 6

Create a dataframe with one row for each of the 20 books, and the book titles, prices, star ratings, and cover JPEG URLs as the four columns. [2 points]

[7]:

```
title_price_ratings_and_url_for_20_books = pd.DataFrame({'title':␣
↪list_of_titles, 'price': list_of_prices, 'rating': list_of_ratings, 'url':␣
↪list_of_urls})
title_price_ratings_and_url_for_20_books
```

[7]:

```
                                                title  price rating  \
0                                A Light in the Attic  51.77  Three
1                                  Tipping the Velvet  53.74    One
2                                          Soumission  50.10    One
3                                       Sharp Objects  47.82   Four
4               Sapiens: A Brief History of Humankind  54.23   Five
5                                     The Requiem Red  22.65    One
6   The Dirty Little Secrets of Getting Your Dream…  33.34   Four
7   The Coming Woman: A Novel Based on the Life of…  17.93  Three
8   The Boys in the Boat: Nine Americans and Their…  22.60   Four
9                                     The Black Maria  52.15    One
10     Starving Hearts (Triangular Trade Trilogy, #1)  13.99    Two
11                              Shakespeare's Sonnets  20.66   Four
12                                        Set Me Free  17.46   Five
13  Scott Pilgrim's Precious Little Life (Scott Pi…  52.29   Five
14                          Rip it Up and Start Again  35.02   Five
15  Our Band Could Be Your Life: Scenes from the A…  57.25  Three
16                                               Olio  23.88    One
```

```
17  Mesaerion: The Best Science Fiction Stories 18…  37.59    One
18                     Libertarianism for Beginners  51.33    Two
19                          It's Only the Himalayas  45.17    Two

                                                url
0   http://books.toscrape.commedia/cache/2c/da/2cd…
1   http://books.toscrape.commedia/cache/26/0c/260…
2   http://books.toscrape.commedia/cache/3e/ef/3ee…
3   http://books.toscrape.commedia/cache/32/51/325…
4   http://books.toscrape.commedia/cache/be/a5/bea…
5   http://books.toscrape.commedia/cache/68/33/683…
6   http://books.toscrape.commedia/cache/92/27/922…
7   http://books.toscrape.commedia/cache/3d/54/3d5…
8   http://books.toscrape.commedia/cache/66/88/668…
9   http://books.toscrape.commedia/cache/58/46/584…
10  http://books.toscrape.commedia/cache/be/f4/bef…
11  http://books.toscrape.commedia/cache/10/48/104…
12  http://books.toscrape.commedia/cache/5b/88/5b8…
13  http://books.toscrape.commedia/cache/94/b1/94b…
14  http://books.toscrape.commedia/cache/81/c4/81c…
15  http://books.toscrape.commedia/cache/54/60/546…
16  http://books.toscrape.commedia/cache/55/33/553…
17  http://books.toscrape.commedia/cache/09/a3/09a…
18  http://books.toscrape.commedia/cache/0b/bc/0bb…
19  http://books.toscrape.commedia/cache/27/a5/27a…
```

1.7 Problem 7

Create a function that takes the URL of the webpage to scrape as an input, applies the code you wrote for questions 1 through 6, and generates the dataframe from question 6 as the output. [3 points]

[8]:

```python
def scrape(url):
    headers = {
        'user-agent': 'Lab Assignment 5: Web Scraping, version 0.1.0␣
 ↪(tsl2b@virginia.edu) (Language=Python 3.9.13; Platform=Mac OSX 13.1)'}
    response = requests.get(url, headers = headers)
    text = response.text
    beautifulSoup = BeautifulSoup(text, 'html')
    list_of_anchors = beautifulSoup.find_all('a', title = True)
    list_of_titles = [a['title'] for a in list_of_anchors]
    list_of_paragraphs = beautifulSoup.find_all('p', 'price_color')
    list_of_prices = [p.string.replace('Â£', '') for p in list_of_paragraphs]
    list_of_paragraphs = beautifulSoup.find_all('p', 'star-rating')
    list_of_ratings = [p['class'][1] for p in list_of_paragraphs]
    list_of_images = beautifulSoup.find_all('img')
    list_of_urls = [url + img['src'] for img in list_of_images]
    title_price_ratings_and_url_for_20_books = pd.DataFrame({'title':␣
↪list_of_titles, 'price': list_of_prices, 'rating': list_of_ratings, 'url':␣
↪list_of_urls})
    return title_price_ratings_and_url_for_20_books

scrape('http://books.toscrape.com')
```

[8]:

```
                                                title  price rating  \
0                                A Light in the Attic  51.77  Three
1                                  Tipping the Velvet  53.74    One
2                                          Soumission  50.10    One
3                                       Sharp Objects  47.82   Four
4               Sapiens: A Brief History of Humankind  54.23   Five
5                                     The Requiem Red  22.65    One
6   The Dirty Little Secrets of Getting Your Dream…  33.34   Four
7   The Coming Woman: A Novel Based on the Life of…  17.93  Three
8   The Boys in the Boat: Nine Americans and Their…  22.60   Four
9                                     The Black Maria  52.15    One
10     Starving Hearts (Triangular Trade Trilogy, #1)  13.99    Two
11                              Shakespeare's Sonnets  20.66   Four
12                                        Set Me Free  17.46   Five
13  Scott Pilgrim's Precious Little Life (Scott Pi…  52.29   Five
14                          Rip it Up and Start Again  35.02   Five
15  Our Band Could Be Your Life: Scenes from the A…  57.25  Three
16                                               Olio  23.88    One
17  Mesaerion: The Best Science Fiction Stories 18…  37.59    One
18                       Libertarianism for Beginners  51.33    Two
19                            It's Only the Himalayas  45.17    Two

                                                  url
0   http://books.toscrape.commedia/cache/2c/da/2cd…
1   http://books.toscrape.commedia/cache/26/0c/260…
2   http://books.toscrape.commedia/cache/3e/ef/3ee…
3   http://books.toscrape.commedia/cache/32/51/325…
4   http://books.toscrape.commedia/cache/be/a5/bea…
5   http://books.toscrape.commedia/cache/68/33/683…
6   http://books.toscrape.commedia/cache/92/27/922…
7   http://books.toscrape.commedia/cache/3d/54/3d5…
8   http://books.toscrape.commedia/cache/66/88/668…
9   http://books.toscrape.commedia/cache/58/46/584…
10  http://books.toscrape.commedia/cache/be/f4/bef…
11  http://books.toscrape.commedia/cache/10/48/104…
12  http://books.toscrape.commedia/cache/5b/88/5b8…
13  http://books.toscrape.commedia/cache/94/b1/94b…
14  http://books.toscrape.commedia/cache/81/c4/81c…
15  http://books.toscrape.commedia/cache/54/60/546…
16  http://books.toscrape.commedia/cache/55/33/553…
```

```
17  http://books.toscrape.commedia/cache/09/a3/09a…
18  http://books.toscrape.commedia/cache/0b/bc/0bb…
19  http://books.toscrape.commedia/cache/27/a5/27a…
```

1.8 Problem 8

Notice that there are many pages to http://books.toscrape.com/. When you click on “Next” in the bottom-right corner of the screen, it takes you to http://books.toscrape.com/catalogue/page-2.html. The front page is the same as http://books.toscrape.com/catalogue/page-1.html, and there are 50 total pages.

Write a loop that uses the function you wrote in question 7 to scrape each of the 50 pages, and append each of these data frames together. If you write this loop correctly, your dataframe will have 1000 rows (20 books on each of the 50 pages).

Some hints:

- Typing `new_df = pd.DataFrame()` with nothing in the parentheses will create an empty data frame on which new data can be appended.
- There are many loops you can use, but the most straightforward one is a for-values loop that counts from 1 to 50. In Python, you can initialize such a loop with for i in range(1, 51):, and indenting every line below it that belongs inside the loop. Inside the loop, the letter i is now a stand-in for the number currently being considered.
- You will need to figure out how to replace the number in URLs like http://books.toscrape.com/catalogue/page-2.html with the number currently under consideration in the loop. You might need the `str()` function, which turns numeric values into strings.

[3 points]

[11]:

```
title_price_ratings_and_url_for_all_books = pd.DataFrame()
for i in range(1, 51):
    url = root + '/catalogue/page-' + str(i) + '.html'
    title_price_ratings_and_url_for_20_books = scrape(url)
    title_price_ratings_and_url_for_all_books =␣
 ↪title_price_ratings_and_url_for_all_books.
 ↪append(title_price_ratings_and_url_for_20_books, ignore_index = True)
print(len(title_price_ratings_and_url_for_all_books.index))
title_price_ratings_and_url_for_all_books.head(n = 41)
```

```
1000
```

[11]:

```
                                          title  price rating  \
0                            A Light in the Attic  51.77  Three
1                              Tipping the Velvet  53.74    One
2                                      Soumission  50.10    One
3                                   Sharp Objects  47.82   Four
4           Sapiens: A Brief History of Humankind  54.23   Five
```

```
5                                     The Requiem Red  22.65     One
6   The Dirty Little Secrets of Getting Your Dream…  33.34    Four
7   The Coming Woman: A Novel Based on the Life of…  17.93   Three
8   The Boys in the Boat: Nine Americans and Their…  22.60    Four
9                                     The Black Maria  52.15     One
10     Starving Hearts (Triangular Trade Trilogy, #1)  13.99     Two
11                              Shakespeare's Sonnets  20.66    Four
12                                        Set Me Free  17.46    Five
13  Scott Pilgrim's Precious Little Life (Scott Pi…  52.29    Five
14                          Rip it Up and Start Again  35.02    Five
15  Our Band Could Be Your Life: Scenes from the A…  57.25   Three
16                                               Olio  23.88     One
17  Mesaerion: The Best Science Fiction Stories 18…  37.59     One
18                       Libertarianism for Beginners  51.33     Two
19                            It's Only the Himalayas  45.17     Two
20                                        In Her Wake  12.84     One
21                                    How Music Works  37.32     Two
22  Foolproof Preserving: A Guide to Small Batch J…  30.52   Three
23                         Chase Me (Paris Nights #2)  25.27    Five
24                                         Black Dust  34.53    Five
25                      Birdsong: A Story in Pictures  54.64   Three
26  America's Cradle of Quarterbacks: Western Penn…  22.50   Three
27                     Aladdin and His Wonderful Lamp  53.13   Three
28  Worlds Elsewhere: Journeys Around Shakespeareâ…  40.30    Five
29                                     Wall and Piece  44.18    Four
30  The Four Agreements: A Practical Guide to Pers…  17.66    Five
31  The Five Love Languages: How to Express Heartf…  31.05   Three
32                                  The Elephant Tree  23.82    Five
33                             The Bear and the Piano  36.89     One
34                                     Sophie's World  15.94    Five
35                                        Penny Maybe  33.29   Three
36     Maude (1883-1993):She Grew Up with the country  18.02     Two
37                               In a Dark, Dark Wood  19.63     One
38                                Behind Closed Doors  52.22    Four
39                     You can't bury them all: Poems  33.63     Two
40                     Slow States of Collapse: Poems  57.31   Three

                                                  url
0   http://books.toscrape.com/catalogue/page-1.htm…
1   http://books.toscrape.com/catalogue/page-1.htm…
2   http://books.toscrape.com/catalogue/page-1.htm…
3   http://books.toscrape.com/catalogue/page-1.htm…
4   http://books.toscrape.com/catalogue/page-1.htm…
5   http://books.toscrape.com/catalogue/page-1.htm…
6   http://books.toscrape.com/catalogue/page-1.htm…
7   http://books.toscrape.com/catalogue/page-1.htm…
8   http://books.toscrape.com/catalogue/page-1.htm…
```

```
9   http://books.toscrape.com/catalogue/page-1.htm…
10  http://books.toscrape.com/catalogue/page-1.htm…
11  http://books.toscrape.com/catalogue/page-1.htm…
12  http://books.toscrape.com/catalogue/page-1.htm…
13  http://books.toscrape.com/catalogue/page-1.htm…
14  http://books.toscrape.com/catalogue/page-1.htm…
15  http://books.toscrape.com/catalogue/page-1.htm…
16  http://books.toscrape.com/catalogue/page-1.htm…
17  http://books.toscrape.com/catalogue/page-1.htm…
18  http://books.toscrape.com/catalogue/page-1.htm…
19  http://books.toscrape.com/catalogue/page-1.htm…
20  http://books.toscrape.com/catalogue/page-2.htm…
21  http://books.toscrape.com/catalogue/page-2.htm…
22  http://books.toscrape.com/catalogue/page-2.htm…
23  http://books.toscrape.com/catalogue/page-2.htm…
24  http://books.toscrape.com/catalogue/page-2.htm…
25  http://books.toscrape.com/catalogue/page-2.htm…
26  http://books.toscrape.com/catalogue/page-2.htm…
27  http://books.toscrape.com/catalogue/page-2.htm…
28  http://books.toscrape.com/catalogue/page-2.htm…
29  http://books.toscrape.com/catalogue/page-2.htm…
30  http://books.toscrape.com/catalogue/page-2.htm…
31  http://books.toscrape.com/catalogue/page-2.htm…
32  http://books.toscrape.com/catalogue/page-2.htm…
33  http://books.toscrape.com/catalogue/page-2.htm…
34  http://books.toscrape.com/catalogue/page-2.htm…
35  http://books.toscrape.com/catalogue/page-2.htm…
36  http://books.toscrape.com/catalogue/page-2.htm…
37  http://books.toscrape.com/catalogue/page-2.htm…
38  http://books.toscrape.com/catalogue/page-2.htm…
39  http://books.toscrape.com/catalogue/page-2.htm…
40  http://books.toscrape.com/catalogue/page-3.htm…
```

[ ]: