Lab Assignment 5: Web Scraping

DS 6001: Practice and Application of Data Science

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 or the book cover JPEG's URL. The dataframe will also 1000 rows, one for each of the 1000 books listed on the 50 pages of this website.

Problem 0

Import the following libraries:

In [1]:

```
import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup as soup
import sys
sys.tracebacklimit = 0 # turn off the error tracebacks
```

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]

Answer 1

In [2]:

```
url = 'http://books.toscrape.com/'
my_headers = {'User-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
r = requests.get(url, headers = my_headers)
r
```

Out[2]:

```
<Response [200]>
```

In [3]:

```
books_scrape = soup(r.text, 'html.parser')
```

Problem 2

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

Answer 2

In [4]:

```
titles = [b.string for b in books_scrape.find_all('h3')]
titles
```

Out[4]:

```
['A Light in the ...',
 'Tipping the Velvet',
 'Soumission',
 'Sharp Objects',
 'Sapiens: A Brief History ...',
 'The Requiem Red',
 'The Dirty Little Secrets ...',
 'The Coming Woman: A ...',
 'The Boys in the ...',
 'The Black Maria',
 'Starving Hearts (Triangular Trade ...',
 "Shakespeare's Sonnets",
 'Set Me Free',
 "Scott Pilgrim's Precious Little ...",
 'Rip it Up and ...',
 'Our Band Could Be ...',
 'Olio',
 'Mesaerion: The Best Science ...',
 'Libertarianism for Beginners',
 "It's Only the Himalayas"]
```

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]

Answer 3

In [5]:

```
prices = [p.string.replace('Â£', '') for p in books_scrape.find_all(class_='price_color')]
prices
```

Out[5]:

```
['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']
```

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]

Answer 4

In [6]:

```
ratings = [r.get('class')[1] for r in books_scrape.find_all(class_='star-rating')]
ratings
```

Out[6]:

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

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]

Answer 5

In [7]:

```
pics = [url+'/'+p.get('src') for p in books_scrape.find_all(class_='thumbnail')]
pics
```

Out[7]:

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

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]

Answer 6

In [8]:

```
df = pd.DataFrame(list(zip(titles, prices, ratings, pics)), columns = ['Title','Price','Rating','P
df
```

Out[8]:

	Title	Price	Rating	Pic
0	A Light in the ...	51.77	Three	http://books.toscrape.com//media/cache/2c/da/2...
1	Tipping the Velvet	53.74	One	http://books.toscrape.com//media/cache/26/0c/2...
2	Soumission	50.10	One	http://books.toscrape.com//media/cache/3e/ef/3...
3	Sharp Objects	47.82	Four	http://books.toscrape.com//media/cache/32/51/3...
4	Sapiens: A Brief History ...	54.23	Five	http://books.toscrape.com//media/cache/be/a5/b...
5	The Requiem Red	22.65	One	http://books.toscrape.com//media/cache/68/33/6...
6	The Dirty Little Secrets ...	33.34	Four	http://books.toscrape.com//media/cache/92/27/9...
7	The Coming Woman: A ...	17.93	Three	http://books.toscrape.com//media/cache/3d/54/3...
8	The Boys in the ...	22.60	Four	http://books.toscrape.com//media/cache/66/88/6...
9	The Black Maria	52.15	One	http://books.toscrape.com//media/cache/58/46/5...
10	Starving Hearts (Triangular Trade ...	13.99	Two	http://books.toscrape.com//media/cache/be/f4/b...
11	Shakespeare's Sonnets	20.66	Four	http://books.toscrape.com//media/cache/10/48/1...
12	Set Me Free	17.46	Five	http://books.toscrape.com//media/cache/5b/88/5...
13	Scott Pilgrim's Precious Little ...	52.29	Five	http://books.toscrape.com//media/cache/94/b1/9...
14	Rip it Up and ...	35.02	Five	http://books.toscrape.com//media/cache/81/c4/8...
15	Our Band Could Be ...	57.25	Three	http://books.toscrape.com//media/cache/54/60/5...
16	Olio	23.88	One	http://books.toscrape.com//media/cache/55/33/5...
17	Mesaerion: The Best Science ...	37.59	One	http://books.toscrape.com//media/cache/09/a3/0...
18	Libertarianism for Beginners	51.33	Two	http://books.toscrape.com//media/cache/0b/bc/0...
19	It's Only the Himalayas	45.17	Two	http://books.toscrape.com//media/cache/27/a5/2...

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]

Answer 7

In [9]:

```
def scrape_page(url):
    r = requests.get(url, headers = {'User-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) Appl
```

```python
    books_scrape = soup(r.text, 'html.parser')
    titles = [b.string for b in books_scrape.find_all('h3')]
    prices = [p.string.replace('Â£', '') for p in books_scrape.find_all(class_='price_color')]
    ratings = [r.get('class')[1] for r in books_scrape.find_all(class_='star-rating')]
    pics = [url+'/'+p.get('src') for p in books_scrape.find_all(class_='thumbnail')]
    return pd.DataFrame(list(zip(titles, prices, ratings, pics)), columns = ['Title','Price','Rati

scrape_page(url)
```

```
Out[9]:
                                Title  Price  Rating                                                Pic
 0                   A Light in the ...  51.77   Three  http://books.toscrape.com//media/cache/2c/da/2...
 1                   Tipping the Velvet  53.74     One  http://books.toscrape.com//media/cache/26/0c/2...
 2                           Soumission  50.10     One  http://books.toscrape.com//media/cache/3e/ef/3...
 3                        Sharp Objects  47.82    Four  http://books.toscrape.com//media/cache/32/51/3...
 4           Sapiens: A Brief History ...  54.23    Five  http://books.toscrape.com//media/cache/be/a5/b...
 5                      The Requiem Red  22.65     One  http://books.toscrape.com//media/cache/68/33/6...
 6         The Dirty Little Secrets ...  33.34    Four  http://books.toscrape.com//media/cache/92/27/9...
 7             The Coming Woman: A ...  17.93   Three  http://books.toscrape.com//media/cache/3d/54/3...
 8                   The Boys in the ...  22.60    Four  http://books.toscrape.com//media/cache/66/88/6...
 9                      The Black Maria  52.15     One  http://books.toscrape.com//media/cache/58/46/5...
10  Starving Hearts (Triangular Trade ...  13.99     Two  http://books.toscrape.com//media/cache/be/f4/b...
11                Shakespeare's Sonnets  20.66    Four  http://books.toscrape.com//media/cache/10/48/1...
12                          Set Me Free  17.46    Five  http://books.toscrape.com//media/cache/5b/88/5...
13         Scott Pilgrim's Precious Little ...  52.29    Five  http://books.toscrape.com//media/cache/94/b1/9...
14                     Rip it Up and ...  35.02    Five  http://books.toscrape.com//media/cache/81/c4/8...
15                 Our Band Could Be ...  57.25   Three  http://books.toscrape.com//media/cache/54/60/5...
16                                 Olio  23.88     One  http://books.toscrape.com//media/cache/55/33/5...
17        Mesaerion: The Best Science ...  37.59     One  http://books.toscrape.com//media/cache/09/a3/0...
18            Libertarianism for Beginners  51.33     Two  http://books.toscrape.com//media/cache/0b/bc/0...
19              It's Only the Himalayas  45.17     Two  http://books.toscrape.com//media/cache/27/a5/2...
```

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]

Answer 8

In [10]:

```
catalogue = pd.DataFrame()
base = 'http://books.toscrape.com/catalogue/page-'
for i in range(1, 51):
    url = base + str(i) + '.html'
    catalogue = catalogue.append(scrape_page(url))
catalogue.reset_index(drop = True, inplace = True)
catalogue
```

Out[10]:

	Title	Price	Rating	Pic
0	A Light in the ...	51.77	Three	http://books.toscrape.com/catalogue/page-1.htm...
1	Tipping the Velvet	53.74	One	http://books.toscrape.com/catalogue/page-1.htm...
2	Soumission	50.10	One	http://books.toscrape.com/catalogue/page-1.htm...
3	Sharp Objects	47.82	Four	http://books.toscrape.com/catalogue/page-1.htm...
4	Sapiens: A Brief History ...	54.23	Five	http://books.toscrape.com/catalogue/page-1.htm...
...	...	...	...	...
995	Alice in Wonderland (Alice's ...	55.53	One	http://books.toscrape.com/catalogue/page-50.ht...
996	Ajin: Demi-Human, Volume 1 ...	57.06	Four	http://books.toscrape.com/catalogue/page-50.ht...
997	A Spy's Devotion (The ...	16.97	Five	http://books.toscrape.com/catalogue/page-50.ht...
998	1st to Die (Women's ...	53.98	One	http://books.toscrape.com/catalogue/page-50.ht...
999	1,000 Places to See ...	26.08	Five	http://books.toscrape.com/catalogue/page-50.ht...

1000 rows × 4 columns