30538 Problem Set 5: Web Scraping

Surya Hardiansyah and Astari Raihanah

2024-11-05

Due 11/9 at 5:00PM Central. Worth 100 points + 10 points extra credit.

Submission Steps (10 pts)

1. This problem set is a paired problem set.
2. Play paper, scissors, rock to determine who goes first. Call that person *Partner 1*.
    - Partner 1 (name and cnet ID): Surya Hardiansyah | sur
    - Partner 2 (name and cnet ID): Astari Raihanah | astari
3. Partner 1 will accept the `ps5` and then share the link it creates with their partner. You can only share it with one partner so you will not be able to change it after your partner has accepted.
4. "This submission is our work alone and complies with the 30538 integrity policy." Add your initials to indicate your agreement: **SH AR**
5. "I have uploaded the names of anyone else other than my partner and I worked with on the problem set **here**" (1 point)
6. Late coins used this pset: **00** Late coins left after submission: **04**
7. Knit your `ps5.qmd` to an PDF file to make `ps5.pdf`,
    - The PDF should not be more than 25 pages. Use `head()` and re-size figures when appropriate.
8. (Partner 1): push `ps5.qmd` and `ps5.pdf` to your github repo.
9. (Partner 1): submit `ps5.pdf` via Gradescope. Add your partner on Gradescope.
10. (Partner 1): tag your submission in Gradescope

```
import pandas as pd
import altair as alt
import time

import warnings
warnings.filterwarnings("ignore")
alt.renderers.enable("png")
```

```
RendererRegistry.enable('png')
```

Step 1: Develop initial scraper and crawler

1. Scraping (PARTNER 1)

```
import requests
from bs4 import BeautifulSoup

# Prepare for Scraping
url = "https://oig.hhs.gov/fraud/enforcement/"
response = requests.get(url)

soup = BeautifulSoup(response.text, "lxml")

# Find all relevant elements within the specified container
containers = soup.find_all("div", class_="usa-card__container")

# Extract data from within the specified containers
titles = [container.find("h2",
 ↪  class_="usa-card__heading").get_text(strip=True) for container in
 ↪  containers]
dates = [container.find("span", class_="text-base-dark
 ↪  padding-right-105").get_text(strip=True) for container in containers]
categories = [container.find("li", class_="display-inline-block usa-tag
 ↪  text-no-lowercase text-base-darkest bg-base-lightest
 ↪  margin-right-1").get_text(strip=True) for container in containers]

# Extract links within the specified containers
link_content = [
    "https://oig.hhs.gov" + a["href"]
    for container in containers
```

```
    for a in container.find_all("a", href=True)
    if "/fraud/enforcement/" in a["href"]]

# Combine extracted data into a DataFrame
data = {
    "Title": titles,
    "Date": dates,
    "Category": categories,
    "Link": link_content[:len(titles)]  # Ensure the number of links matches
    ↪ the titles
}
df = pd.DataFrame(data)

# Print the head of df
print(df.head())
```

```
                                               Title              Date  \
0  North Texas Medical Center Pays $14.2 Million ...  November 4, 2024
1  New England Doctor Pleads Guilty To Drug Distr...  November 4, 2024
2  St. Louis County Woman Accused Of $3 Million H...  November 1, 2024
3  Lab Owner And Marketing Company Owner Both Fou...  November 1, 2024
4  Compound Ingredient Supplier Medisca Inc., To ...  November 1, 2024

                    Category  \
0  Criminal and Civil Actions
1  Criminal and Civil Actions
2  Criminal and Civil Actions
3  Criminal and Civil Actions
4  Criminal and Civil Actions

                                                Link
0  https://oig.hhs.gov/fraud/enforcement/north-te...
1  https://oig.hhs.gov/fraud/enforcement/new-engl...
2  https://oig.hhs.gov/fraud/enforcement/st-louis...
3  https://oig.hhs.gov/fraud/enforcement/lab-owne...
4  https://oig.hhs.gov/fraud/enforcement/compound...
```

2. Crawling (PARTNER 1)

```python
# Function to extract agency information from a given link
def extract_agency_name(link):
    response = requests.get(link)
    soup = BeautifulSoup(response.text, 'lxml')

    # Find the div containing the action details
    details_div = soup.find('div', class_="margin-top-5 padding-y-3
 ↪  border-y-1px border-base-lighter")
    if details_div:
        # Find all <li> elements within the div
        li_elements = details_div.find_all('li')
        for li in li_elements:
            # Check if the <li> contains 'Agency:' and extract the text after
             ↪  it
            if 'Agency:' in li.get_text():
                return li.get_text().replace('Agency:', '').strip()

    return 'N/A'

# Append agency info into df
df["Agency"] = df["Link"].apply(extract_agency_name)

# Print the head of df
print(df.head())
```

```
                                               Title              Date  \
0  North Texas Medical Center Pays $14.2 Million ...  November 4, 2024
1  New England Doctor Pleads Guilty To Drug Distr...  November 4, 2024
2  St. Louis County Woman Accused Of $3 Million H...  November 1, 2024
3  Lab Owner And Marketing Company Owner Both Fou...  November 1, 2024
4  Compound Ingredient Supplier Medisca Inc., To ...  November 1, 2024

                    Category  \
0  Criminal and Civil Actions
1  Criminal and Civil Actions
2  Criminal and Civil Actions
3  Criminal and Civil Actions
4  Criminal and Civil Actions

                                                Link  \
0  https://oig.hhs.gov/fraud/enforcement/north-te...
```

```
1  https://oig.hhs.gov/fraud/enforcement/new-engl...
2  https://oig.hhs.gov/fraud/enforcement/st-louis...
3  https://oig.hhs.gov/fraud/enforcement/lab-owne...
4  https://oig.hhs.gov/fraud/enforcement/compound...

                                                Agency
0  U.S. Attorney's Office, Northern District of T...
1                          U.S. Department of Justice
2  U.S. Attorney's Office, Eastern District of Mi...
3  U.S. Attorney's Office, Middle District of Ten...
4                          U.S. Department of Justice
```

Step 2: Making the scraper dynamic

1. Turning the scraper into a function

- a. Pseudo-Code (PARTNER 2)
- b. Create Dynamic Scraper (PARTNER 2)
- c. Test Partner"s Code (PARTNER 1)

Step 3: Plot data based on scraped data

1. Plot the number of enforcement actions over time (PARTNER 2)

2. Plot the number of enforcement actions categorized: (PARTNER 1)

- based on "Criminal and Civil Actions" vs. "State Enforcement Agencies"
- based on five topics

Step 4: Create maps of enforcement activity

1. Map by State (PARTNER 1)

2. Map by District (PARTNER 2)

Extra Credit

1. Merge zip code shapefile with population

2. Conduct spatial join

3. Map the action ratio in each district