Pset V

Summer Negahdar & Jenny Zhong

Partner 1: Summer Negahdar(samarneg) Partner 2: This submission is our work alone and complies with the 30538 integrity policy." Add your initials to indicate your agreement: **___** **"I have uploaded the names of anyone else other than my partner and I worked with on the problem set here" Late coins used this pset:** Late coins left after submission: **___**

```
from bs4 import BeautifulSoup
import pandas as pd
import requests
```

Develop Initial scraper and crawler

1.

```
#doing the intial action to get the link parsed
url = 'https://oig.hhs.gov/fraud/enforcement/' # the link we would be scraping
requested = requests.get(url)
soup = BeautifulSoup(requested.content, 'html.parser')

# Find all actions based on the main <li> tag containing each card
actions = soup.find_all('li', class_='usa-card card--list pep-card--minimal
↪  mobile:grid-col-12')

dataset = [] #creating an ampty list to store craped data

for items in actions:
    title_tag = items.find('h2', class_='usa-card__heading') #tag for the title of
↪  enforcement is h2
    if title_tag:
        title = title_tag.get_text(strip=True)
        link = title_tag.find('a')['href'] if title_tag.find('a') else None #the tag for
↪  hyperlinks
        link = f"https://oig.hhs.gov{link}" if link else None  # Complete relative link

    #looking for dates
    date_tag = items.find(lambda tag: tag.name == "span" and "text-base-dark" in
↪  tag.get("class", []) and "padding-right-105" in tag.get("class", []))
    date = date_tag.get_text(strip=True) if date_tag else None
```

```
    #now we will be looking for category
    category_tag = items.find(lambda tag: tag.name == "ul" and "display-inline" in
↪   tag.get("class", []) and "add-list-reset" in tag.get("class", []))
    category = None
    if category_tag:
        li_tag = category_tag.find(lambda tag: tag.name == "li" and
↪   "display-inline-block" in tag.get("class", []) and "usa-tag" in tag.get("class", []))
        category = li_tag.get_text(strip=True) if li_tag else None

    # Append data to dataset
    dataset.append({
        'Title': title,
        'Date': date,
        'Category': category,
        'Link': link
    })

final_df = pd.DataFrame(dataset)
print(final_df.head(5))
```

```
                                               Title              Date  \
0  Pharmacist and Brother Convicted of $15M Medic...  November 8, 2024
1  Boise Nurse Practitioner Sentenced To 48 Month...  November 7, 2024
2  Former Traveling Nurse Pleads Guilty To Tamper...  November 7, 2024
3  Former Arlington Resident Sentenced To Prison ...  November 7, 2024
4  Paroled Felon Sentenced To Six Years For Fraud...  November 7, 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/pharmaci...
1  https://oig.hhs.gov/fraud/enforcement/boise-nu...
2  https://oig.hhs.gov/fraud/enforcement/former-t...
3  https://oig.hhs.gov/fraud/enforcement/former-a...
4  https://oig.hhs.gov/fraud/enforcement/paroled-...
```

2.

```
# Initialize an empty list to store agency names
agencies = []

# Loop through each link in final_df
```

```
for index, row in final_df.iterrows():
    link = row['Link']
    if link:  # Only proceed if the link is valid
        try:
            response = requests.get(link)  # Request the page using the link
            soup = BeautifulSoup(response.text, 'html.parser')  # Parse the content of
 ↪  the page

            # Find all <ul> elements with the specified class containing the agency
             ↪  details
            uls = soup.find_all("ul", class_="usa-list usa-list--unstyled margin-y-2")

            # Initialize a placeholder for the agency name
            agency_name = 'N/A'

            # Iterate over each <ul> element
            for ul in uls:
                # Find all <span> elements within each <ul> that match the class
                spans = ul.find_all("span", class_="padding-right-2 text-base")

                # Ensure there are enough <span> tags to avoid IndexError
                if len(spans) > 1:
                    agency = spans[1]  # Select the second <span>, which contains
 ↪  "Agency:"

                    # Use next_sibling to access the text following the <span>
                    agency_name = agency.next_sibling.strip() if agency.next_sibling else
 ↪  'N/A'

                    # Stop after finding the first matching <ul> and <span> structure
                    break

            # Append the extracted agency name to the agencies list
            agencies.append(agency_name)

        except requests.exceptions.RequestException as e:
            print(f"Error fetching {link}: {e}")
            agencies.append('N/A')

# Add agency names to the DataFrame and print its head
final_df['Agency'] = agencies  # Create a new column in our original df called Agency
print(final_df.head(5))
```

```
                                               Title             Date  \
0  Pharmacist and Brother Convicted of $15M Medic...  November 8, 2024
1  Boise Nurse Practitioner Sentenced To 48 Month...  November 7, 2024
2  Former Traveling Nurse Pleads Guilty To Tamper...  November 7, 2024
3  Former Arlington Resident Sentenced To Prison ...  November 7, 2024
4  Paroled Felon Sentenced To Six Years For Fraud...  November 7, 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/pharmaci...
1  https://oig.hhs.gov/fraud/enforcement/boise-nu...
2  https://oig.hhs.gov/fraud/enforcement/former-t...
3  https://oig.hhs.gov/fraud/enforcement/former-a...
4  https://oig.hhs.gov/fraud/enforcement/paroled-...

                                             Agency
0                          U.S. Department of Justice
1  November 7, 2024; U.S. Attorney's Office, Dist...
2  U.S. Attorney's Office, District of Massachusetts
3  U.S. Attorney's Office, Eastern District of Vi...
4  U.S. Attorney's Office, Middle District of Flo...
```

```python
##for jenny
## since you need the date column to be a date, I will convert it for you

final_df['Date'] = pd.to_datetime(final_df['Date'], errors='coerce')

# Check the data type to confirm
print(final_df.dtypes)
print(final_df.head())
```

```
Title               object
Date        datetime64[ns]
Category            object
Link                object
Agency              object
dtype: object
                                               Title       Date  \
0  Pharmacist and Brother Convicted of $15M Medic... 2024-11-08
1  Boise Nurse Practitioner Sentenced To 48 Month... 2024-11-07
2  Former Traveling Nurse Pleads Guilty To Tamper... 2024-11-07
3  Former Arlington Resident Sentenced To Prison ... 2024-11-07
4  Paroled Felon Sentenced To Six Years For Fraud... 2024-11-07

                    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/pharmaci...
1  https://oig.hhs.gov/fraud/enforcement/boise-nu...
2  https://oig.hhs.gov/fraud/enforcement/former-t...
3  https://oig.hhs.gov/fraud/enforcement/former-a...
4  https://oig.hhs.gov/fraud/enforcement/paroled-...

                                              Agency
0                          U.S. Department of Justice
1  November 7, 2024; U.S. Attorney's Office, Dist...
2  U.S. Attorney's Office, District of Massachusetts
3  U.S. Attorney's Office, Eastern District of Vi...
4  U.S. Attorney's Office, Middle District of Flo...
```

Making the scraper dynamic

1.

the pseudo code for writing the function will be like: 1. going thorugh every row in the df Summer has created.

2. extracting the dates from date column
3. there will be two types of date

   I. after 2013 » continue with the rest of function

   II. before 2013 » show me an error sign that says this is before our desired timeline

4. save all the extracted dates on a csv file called “enfrocment_actions_month_year.csv”
5. do not push it to git
6. print the head

a.
b.
c.

Plot data based on scraped data (using Altair)

1.

2.

Create maps of enforcement activity

1.

2.

Extra Credit: Calculate the enforcement actions on a per-capita basis

1.

2.

3.