Problem Set 5

Lauren Laine and Mohamed Mohamed

Invalid Date

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): Lauren Laine, llaine
 - Partner 2 (name and cnet ID):
- 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: **__** **__**
- 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: ** ** Late coins left after submission: ** **
- 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
url = 'https://oig.hhs.gov/fraud/enforcement/'
response = requests.get(url)
soup = BeautifulSoup(response.text, 'lxml')
soup.text[0:50]
```

Boise Health Care Company And Former Therapist Resolve Fraudulent Psychotherapy Billing Allegations Involving Vulnerable Refugee Population

```
#scrape date
dates=[]
soup_dates=soup.find_all('span', class_='text-base-dark padding-right-105')
for tag in soup_dates:
  text=tag.text
  dates.append(text)
print(dates[0:5])
print(dates[19])
['November 7, 2024', 'November 7, 2024', 'November 7, 2024', 'November 7,
2024', 'November 6, 2024']
October 29, 2024
# scrape category
category=[]
soup_category=soup.find_all('li', class_="display-inline-block usa-tag

→ text-no-lowercase text-base-darkest bg-base-lightest margin-right-1")

for tag in soup_category:
 text=tag.text
  category.append(text)
print(category[19])
```

Criminal and Civil Actions

```
#scrape link associated with the enforecment action
hrefs=[]
link_tags=[]
full_links=[]
for tag in usa_card_heading:
    link_tags.append(tag.find('a').attrs)

for link in link_tags:
    href=link.get('href')
    hrefs.append(href)

print(hrefs[19])
prefix='https://oig.hhs.gov/'
for href in hrefs:
    link= prefix+href
    full_links.append(link)
```