MSADS509 Group 3 Final Project: Data Pulling

Importing Libraries

In [1]:

```
import datetime
import random
import requests
import time
from bs4 import BeautifulSoup
from urllib.parse import urljoin
from collections import defaultdict
import pandas as pd
```

Scraping Political Data from CNN and Fox News

Below we create a function that scrapes the articles of the day for both Fox News and CNN and creates a pandas data frame using the content that is pulled from the articles. This function pulls the daily articles, and we ran it every day for five consecutive weekdays to get a full business week's worth of data for our topic modeling. This function is here for purposes of showing our methods, but we will ultimately construct the data frame to be cleaned in the cell beneath it by concatenating the five CSV files that were pulled.

In [2]:

```
def return_text_if_not_none(element):
    return element.get_text(separator=' ', strip=True) if element else None

current_year = datetime.datetime.now().year

source = {'cnn': "https://www.cnn.com/politics",
          'foxnews': "https://www.foxnews.com/politics"}

news_pages = defaultdict(list)  # Use a list to store URLs and content

for source_name, source_page in source.items():

    # request the page and sleep
    r = requests.get(source_page)

    time.sleep(5 + 10 * random.random())

    soup = BeautifulSoup(r.content, 'html.parser')

    links = soup.find_all('a', href=True)

    for link in links:
        href = link['href']
        # Convert relative URLs to absolute URLs
        full_url = urljoin(source_page, href)

        # Check if the link contains "/politics/" and does not contain "/gal
        if "/politics/" in full_url and "/gallery/" not in full_url:

            # Check if it's CNN and the URL has the format 'cnn.com/{}/'
            if source_name == 'cnn' and f"cnn.com/{current_year}/" in full_u

                # Fetch the news content
                content_r = requests.get(full_url)

                content_soup = BeautifulSoup(content_r.content, 'html.parser

                article_content = return_text_if_not_none(content_soup.find(

                news_pages[source_name].append({'url': full_url, 'content':

            # Check if it's FOXNEWS and the URL does not contain "/category/
            elif source_name == 'foxnews' and "/category/" not in full_url:

                # Fetch the news content
                content_r = requests.get(full_url)

                content_soup = BeautifulSoup(content_r.content, 'html.parser

                article_content = return_text_if_not_none(content_soup.find(

                news_pages[source_name].append({'url': full_url, 'content':
# Create a DataFrame

df = pd.DataFrame([(source_name, item['url'], item['content']) for source_na
                   news_pages.items() for item in items], columns=['source',

df = df.drop_duplicates()

df.head()
```

Out[2]:

	source	url	content
0	cnn	https://www.cnn.com/2024/02/16/politics/russia...	CNN — Russia is trying to develop a nuclear sp...
2	cnn	https://www.cnn.com/2024/02/15/politics/takeaw...	CNN — The Georgia election subversion case aga...
3	cnn	https://www.cnn.com/2024/02/16/politics/biden-...	Washington CNN — The Norfolk Southern train de...
4	cnn	https://www.cnn.com/2024/02/16/politics/gaetz-...	CNN — The House Ethics Committee investigating...
5	cnn	https://www.cnn.com/2024/02/16/politics/takeaw...	CNN — Judge Arthur Engoron hit Donald Trump wi...

News Counts for CNN and Fox News

In [3]:

```
source_counts = df['source'].value_counts()

# Print the counts for each source
print("CNN rows:", source_counts.get('cnn', 0))
print("Fox News rows:", source_counts.get('foxnews', 0))
```

```
CNN rows: 48
Fox News rows: 20
```

Saving FoxNews and CNN Political News Results from Feb 12 to Feb 16 to Local Storage

In [4]:

```
#df.to_csv('/Users/UE/Desktop/MSADS509_News_Project_Dataset/news_0212.csv',
#df.to_csv('/Users/UE/Desktop/MSADS509_News_Project_Dataset/news_0213.csv',
#df.to_csv('/Users/UE/Desktop/MSADS509_News_Project_Dataset/news_0214.csv',
#df.to_csv('/Users/UE/Desktop/MSADS509_News_Project_Dataset/news_0215.csv',
df.to_csv('/Users/UE/Desktop/MSADS509_News_Project_Dataset/news_0216.csv', i
```