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
```

In [6]:

```
import pandas as pd
import re

from nltk.corpus import stopwords
from string import punctuation

from collections import Counter, defaultdict

import glob
from collections import Counter

from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
import nltk
nltk.download('punkt')

import warnings
warnings.filterwarnings("ignore")
```

```
[nltk_data] Downloading package punkt to /Users/kevinbaum/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

We start by accessing the five files that were pulled during the week on Monday 02/12/2024 and combining them into a DataFrame.

In [7]:

```
# File paths
files = ['Data/news_0212.csv', 'Data/news_0213.csv',
         'Data/news_0214.csv', 'Data/news_0215.csv',
        'Data/news_0216.csv']

# Read each CSV file into a DataFrame and store them in a list
file_dfs = [pd.read_csv(file) for file in files]

# Concatenate all DataFrames into one
combined_df = pd.concat(file_dfs, ignore_index=True)
combined_df
```

Out[7]:

	source	url	content
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	CNN — Chairman of the Joint Chiefs of Staff Ge...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	CNN — Former President Donald Trump has endors...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Washington CNN — President Joe Biden and King ...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	CNN — Former President Donald Trump on Monday ...
...	...	...	...
348	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News' Politics newsletter with ...
349	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News' Politics newsletter with ...
350	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News' Politics newsletter with ...
351	foxnews	https://www.foxnews.com/politics/democrats-win...	close Video Dems flipping NY House seat threat...
352	foxnews	https://www.foxnews.com/politics/trumps-nato-c...	close Video The media doesn't allow the public...

353 rows × 3 columns

Before proceeding, we remove the duplicates in the DataFrame.

In [8]:

```
df = combined_df.drop_duplicates().reset_index(drop=True)
df
```

Out[8]:

	source	url	content
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	CNN — Chairman of the Joint Chiefs of Staff Ge...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	CNN — Former President Donald Trump has endors...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Washington CNN — President Joe Biden and King ...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	CNN — Former President Donald Trump on Monday ...
...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republi...	close Video Rep. Ronny Jackson demands Biden t...
236	foxnews	https://www.foxnews.com/politics/gop-senators-...	close Video Biden and the Democrats just do no...
237	foxnews	https://www.foxnews.com/politics/doj-defends-s...	close Video Former US attorney discusses Speci...
238	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News’ Politics newsletter with ...
239	foxnews	https://www.foxnews.com/politics/democrats-win...	close Video Dems flipping NY House seat threat...

240 rows × 3 columns

Checking the Results of the Web Scraping

Confirming that the CNN content was scraped successfully

In [9]:

```
df[df['source']=='cnn'].head()
```

Out[9]:

	source	url	content
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	CNN — Chairman of the Joint Chiefs of Staff Ge...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	CNN — Former President Donald Trump has endors...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Washington CNN — President Joe Biden and King ...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	CNN — Former President Donald Trump on Monday ...

Confirming that the Fox News content was scraped successfully

In [10]:

```
df[df['source']=='foxnews'].head()
```

Out[10]:

	source	url	content
47	foxnews	https://www.foxnews.com/politics/biden-takes-j...	close Video Biden takes jab at special counsel...
48	foxnews	https://www.foxnews.com/politics/rfk-jr-apolog...	close Video RFK Jr. drops surprise campaign ad...
49	foxnews	https://www.foxnews.com/politics/bidens-upcomi...	close Video Biden won't take cognitive test in...
50	foxnews	https://www.foxnews.com/politics/kamala-harris...	close Video Marc Thiessen questions whether Bi...
51	foxnews	https://www.foxnews.com/politics/climate-activ...	close Video Biden's export suspension on lique...

Let's take a full look at one of the rows for both CNN and Fox to see if there are any obvious steps that stand out that we want to clean up.

In [11]:

```
# Set pandas to display the full content of a column
# We will do this only temporarily and reset it after
# testing one column
pd.set_option('display.max_colwidth', None)
```

In [12]:

```
# Display one row from CNN
print("CNN Article Content:")
print(df[df['source'] == 'cnn'].iloc[0])
```

```
CNN Article Content:
source
cnn
url
https://www.cnn.com/2024/02/12/politics/cq-brown-nato-trump/index.html
content    CNN — Chairman of the Joint Chiefs of Staff Gen. CQ Brown said Mo
nday that “US credibility is at stake” in the wake of comments from former P
resident Donald Trump that he would encourage Russia to “do whatever the hel
l they want” to NATO partners that don’t meet spending guidelines on defens
e. Asked by NBC News about Trump’s admission that he would not abide by the
collective-defense clause at the heart of NATO if reelected, Brown said that
the alliance is strong and has been around for 75 years. “I think we have a
responsibility to uphold those alliances,” Brown told NBC’s Lester Holt in a
n interview airing Monday evening. “US credibility is at stake with each of
our alliances, and US leadership is still needed, wanted, and watched.” “The
US is committed,” Brown added. “And that’s the message I communicate, and th
at’s the message that’s been received.” Brown’s remarks come as Trump, the 2
024 Republican front-runner, has come under fire for his comments over the w
eekend indicating he does not intend to defend NATO allies from Russian atta
ck if he is reelected. Former President Donald Trump speaks as he holds a ca
mpaign rally at Coastal Carolina University ahead of the South Carolina Repu
blican presidential primary in Conway, South Carolina, on February 10. Sam W
olfe/Reuters Related article Trump will pull US out of NATO if he wins elect
ion, ex-adviser warns At a rally in South Carolina Saturday, Trump recalled
a conversation he had while president with “one of the presidents of a big c
ountry,” who asked him whether the US would defend them from a Russian invas
ion even if they “don’t pay.” “No, I would not protect you,” he recalled say
ing. “In fact, I would encourage them to do whatever the hell they want. You
got to pay. You got to pay your bills.” According to NBC , Brown said that h
e realizes there will be “various dialogue in discussions at the political l
evel,” but that he will focus on “continuing to build and strengthen our rel
ationship with NATO.” “My job is to make sure that we are doing everything w
e can with our NATO allies on the military aspect,” he said, “and I’ll conti
nue to do that throughout.” Brown, who was sworn in as the chairman of the J
oint Chiefs of staff last fall, serves as the principal military advisor to
the president. The military officer who serves in the role does so at the pl
easure of the president, meaning that if Trump were reelected this year, Bro
wn could serve as his chairman unless he appointed another officer. Gen. Mar
k Milley , Brown’s predecessor and Trump’s chairman of the Joint Chiefs thro
ughout his tenure, was a frequent target of Trump’s ire, and the two have ta
ken verbal – sometimes subtle – shots at each other after Trump left the Whi
te House. In late September, at the change of command ceremony between Mille
y and Brown, the outgoing chairman made it a point to underscore where his l
oyalty and the loyalty of the military should lie. “We don’t take an oath to
a king, or a queen, or a tyrant or a dictator. And we don’t take an oath to
a wannabe dictator,” Milley said at the time, in a veiled reference to Trum
p. “We take an oath to the Constitution and we take an oath to the idea that
is America — and we’re willing to die to protect it.” Milley chose to delive
r the scathing criticism of his former boss in his last address as the natio
n’s top general as he stood next to President Joe Biden. In a continuation o
f the acrimonious feud between the two, Trump fired back on social media, ca
lling Milley a “moron” and “STUPID & VERY DANGEROUS!” Brown addressed recent
criticism of his boss Monday following mention of apparent memory lapses in
special counsel Robert Hur’s report , calling Biden “pretty sharp.” Asked wh
ether he was surprised by the comments made about the president’s memory, Br
own said he was, adding they were “not characteristic of what I’ve seen.” “H
```

```
e’s got a very good grasp of the issues,” Brown said. “He asks, I think, ver
y pertinent questions.” Iran likely ‘not looking for a broader conflict’ wit
h US As the US is navigating increasingly high tensions in the Middle East
– from Iran-backed groups in Iraq and Syria, the Iran-backed Houthis in Yeme
n, and Israel’s campaign in Gaza against Hamas – Brown echoed comments made
by other US officials that deterrence is key. Broadly, the US is focused on
deterring “any further aggression,” Brown said, while also protecting Americ
an forces. Asked whether he believed Iran wants a war with the US, Brown res
ponded, “I don’t know that they do.” “Having watched Iran operate, they will
do things through their militia groups and others to put pressure, to achiev
e their objectives,” the chairman said. “At the same time, not looking for a
broader conflict with the United States.” There have been at least 170 attac
ks on US and coalition forces in Iraq, Syria and Jordan since October 17. Th
e Pentagon said Monday that those attacks have resulted in 186 wounded or ki
lled in action – including 130 traumatic brain injuries. Three US soldiers w
ere killed in a drone attack in January on a US outpost in Jordan. This stor
y has been updated with additional information.
Name: 0, dtype: object
```

In [13]:

```
# Display one row from CNN
print("Fox Article Content:")
print(df[df['source'] == 'foxnews'].iloc[0])
```

```
Fox Article Content:
source
foxnews
url
https://www.foxnews.com/politics/biden-takes-jab-hur-report-joke-memory-retu
rns-speech-one-more-thing-forgot
content    close Video Biden takes jab at special counsel report with joke a
bout his memory President Biden on Monday joked about his memory and age dur
ing a speech to the National Association of Counties in Washington, D.C. Joi
n Fox News for access to this content Plus get unlimited access to thousands
of articles, videos and more with your free account! Please enter a valid em
ail address. By entering your email, you are agreeing to Fox News Terms of S
ervice and Privacy Policy , which includes our Notice of Financial Incentive
. To access the content, check your email and follow the instructions provid
ed. President Biden attempted a joke about his memory during a speech in Was
hington, D.C., Monday, seemingly taking a jab at Special Counsel Robert Hu
r's report. Delivering remarks at the National Association of Counties Legis
lative Conference, Biden spoke about his bipartisan infrastructure law, whic
h he credits for allowing his administration to continue "making the biggest
investment in climate change ever anywhere in the entire world." "After deva
stating floods, tornadoes, wildfires and hurricane, we're going to keep work
ing together to respond, to rebuild and boost resilience to extreme weather.
My administration is also helping install rooftop solar to build a national
network of electric vehicle charging stations for revitalizing fenceline com
munities smothered by the legacy of pollution like where I lived in Claymon
t," Biden said, referring to where his family moved in Delaware during the e
arly 1950s. "We're promoting clean energy in industries of the future made h
ere in America. Made in America," he said while transitioning. BIDEN ALLIES
GO ON DEFENSE BLITZ FOLLOWING HUR REPORT: 'BUCKET OF BS' "What I didn't real
ize, and I've been around, I know it don't look like it, but I've been aroun
d a while. I do remember that," Biden said, garnering laughter and applause.
President Biden gave a speech to the National Association of Counties Legisl
ative Conference, Monday, Feb. 12, 2024, in Washington. (AP Photo/Evan Vucc
i) The joke was somewhat undercut by Biden concluding the speech — and then
adding an interjection at the end, admitting, "I forgot something," before m
aking a final comment. In building his argument for why no charges were reco
mmended following an investigation into Biden's mishandling of classified do
cuments, Hur, who was appointed by Attorney General Merrick Garland, detaile
d in part that Biden's defense of any potential charges could possibly be th
at, "Mr. Biden would likely present himself to a jury, as he did during our
interview of him, as a sympathetic, well-meaning, elderly man with a poor me
mory." The report cited examples when investigators said the president's mem
ory lapsed, including over when his older son Beau had died. Biden's age and
mental fitness have already been a concern for voters. President Biden after
delivering remarks to the National Association of Counties Legislative Confe
rence, Monday, Feb. 12, 2024, in Washington. (AP Photo/Evan Vucci) MAYORKAS
DUCKS RESPONSIBILITY ON BORDER CRISIS, MIGRANT FIGURES: 'CONGRESS IS THE ONL
Y ONE WHO CAN FIX THIS' During his speech, Biden also criticized his 2024 ri
val, former President Trump, and Republicans for opposing a $118 billion sup
plemental spending agreement that included aid for Ukraine, Israel and Taiwa
n, as well as an ambitious border security and immigration package. The bord
er package drew widespread opposition from conservative Republicans in both
chambers since its release just days earlier. The Senate voted against the s
upplemental 50-49 Wednesday. It needed 60 votes to pass. The vote went mostl
y along party lines, except for five Democrats voting no and four Republican
s voting yes. "Some of my extreme Republican friends — and by the way, this
```

```
is not your father's Republican Party … I'm not taking on all Republicans. I
really mean it. The MAGA Republicans, a minority, but a powerful minority. T
hey went out and they killed the deal. My predecessor said he didn't like i
t. It was a loss for him. We have to end the political games, folks," Biden
said Monday. President Biden, left, greets NatCo president and commissioner
in Ramsey County, Minnesota, Mary Jo McGuire, before he delivers remarks to
the National Association of Counties Legislative Conference, Monday, Feb. 1
2, 2024. (AP Photo/Evan Vucci) The president also claimed a victory for the
economy. "It's clear we have the strongest economy in the world. Nearly 15 m
illion new jobs since I came to office," Biden said. "The longest stretch of
under 4% in 50 years. Growth is strong. Rising wages are rising, inflation i
s down. In fact, the costs have fallen from everything from a gallon of gas
to a gallon of milk. We know prices are still too high because of what I cal
l greed-inflation and shrinkflation," he said, referring to companies chargi
ng the same amount for a product while reducing quantity. "I'm calling on co
rporations to pass their savings on to consumers, for God's sake. We're maki
ng real progress." CLICK TO GET THE FOX NEWS APP "The recent Washington Post
headline summed it up," Biden added, quoting the newspaper's story titled,
"Falling Inflation and Rising Growth Give the U.S. the World's Best Recover
y." "The world's best recovery!" Biden said. "It's because you implemented w
hat we did. You made it work." Danielle Wallace is a reporter for Fox News D
igital covering politics, crime, police and more. Story tips can be sent to
danielle.wallace@fox.com and on Twitter: @danimwallace.
Name: 47, dtype: object
```

In [14]:
```python
# Reset the columns so that we go back to truncating the "content" column
pd.reset_option('display.max_colwidth')
```

Data Cleaning, Tokenizing, and Normalizing

Removing Unwanted Prefixes

We see from looking at the first 5 rows of the CNN and Fox records that the content of the articles starts with "CNN --" or "(city name) CNN" for CNN and "close Video" for Fox. Since this is noise in our attempt to topic model, we will remove this part of the content body using the function below.

In [15]:
```python
# Function to remove prefixes

def remove_prefix(row):
    # Pattern to match "CNN —" if it includes a city's name before it
    cnn_pattern = r'^(?:[\w\s]+\s)?CNN — '

    # For CNN, remove pattern if it matches
    if row['source'] == 'cnn':
        return re.sub(cnn_pattern, '', row['content'])

    # For Fox News, remove "close Video " prefix
    elif row['source'] == 'foxnews' and row['content'].startswith('close Vic
        return row['content'][12:]

    # Return original content if no prefix to remove
    else:
        return row['content']

df['content'] = df.apply(remove_prefix, axis=1)
```

In [16]:

```
# Checking to see how CNN looks without the prefixes
df[df['source']=='cnn']
```

Out[16]:

	source	url	content
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	Chairman of the Joint Chiefs of Staff Gen. CQ ...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Former President Donald Trump has endorsed Nor...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	President Joe Biden and King Abdullah II of Jo...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Former President Donald Trump on Monday asked ...
...	...	...	...
220	cnn	https://www.cnn.com/2024/02/15/politics/navy-f...	Members of Congress pressed the CEO of the nat...
221	cnn	https://www.cnn.com/2024/02/16/politics/kamala...	US Vice President Kamala Harris on Friday call...
222	cnn	https://www.cnn.com/2024/01/30/politics/trump-...	New York state Judge Arthur Engoron has the fu...
223	cnn	https://www.cnn.com/2024/02/15/politics/border...	The acting deputy chief of the US Border Patro...
224	cnn	https://www.cnn.com/2024/02/15/politics/former...	Special counsel David Weiss charged a former F...

143 rows × 3 columns

In [17]:

```
# Checking to see how Fox looks without the prefixes
df[df['source']=='foxnews']
```

Out[17]:

	source	url	content
47	foxnews	https://www.foxnews.com/politics/biden-takes-j...	Biden takes jab at special counsel report with...
48	foxnews	https://www.foxnews.com/politics/rfk-jr-apolog...	RFK Jr. drops surprise campaign ad during Supe...
49	foxnews	https://www.foxnews.com/politics/bidens-upcomi...	Biden won't take cognitive test in physical ex...
50	foxnews	https://www.foxnews.com/politics/kamala-harris...	Marc Thiessen questions whether Biden is capab...
51	foxnews	https://www.foxnews.com/politics/climate-activ...	Biden’s export suspension on liquefied natural...
...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republi...	Rep. Ronny Jackson demands Biden take cognitiv...
236	foxnews	https://www.foxnews.com/politics/gop-senators-...	Biden and the Democrats just do not care: Sen....
237	foxnews	https://www.foxnews.com/politics/doj-defends-s...	Former US attorney discusses Special Counsel H...
238	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News’ Politics newsletter with ...
239	foxnews	https://www.foxnews.com/politics/democrats-win...	Dems flipping NY House seat threatens GOP majo...

97 rows × 3 columns

Remove Unwanted First Sentences

We see that some of the CNN articles begin with the following sentences: "A version of this story appeared in CNN’s What Matters newsletter. To get it in your inbox, sign up for free here." We also see that some of the Fox articles begin with the phrase "Welcome to Fox News" in the first sentence. To remove this noise, we will write a function below that handles it. We need to run this function twice in order to fully clean out the noise. Also, once we run the function twice, we need to re-run the "remove_prefix" function again as the prefixes will be present after removing some of the unwanted first sentences.

In [18]:

```
def remove_first_sentence(row):
    # Split the content into sentences based on '.', '?', and '!'
    sentences = re.split(r'(?<=[.!?]) +', row['content'])

    # Initialize updated_content with the original content in case none of t
    updated_content = row['content']

    if len(sentences) > 1:  # Check if there's more than one sentence
        first_sentence = sentences[0]  # Get the first sentence

        if 'Welcome to Fox News' in first_sentence:
            # Join all sentences except the first one. We start at 2 because
            updated_content = ' '.join(sentences[2:])

        elif 'A version of this story appeared' in first_sentence:
            # Removing 2 sentences since CNN includes 2 unwanted sentences i
            updated_content = ' '.join(sentences[2:])

        elif first_sentence.strip().startswith("What's Happening?"):
            # Directly check if the first sentence is exactly "What's Happen
            updated_content = ' '.join(sentences[1:])

    # Remove sentences containing the phrase 'CLICK HERE TO GET THE FOX NEWS
    updated_sentences = [sentence for sentence in sentences if 'FOX NEWS APP
    updated_sentences = [sentence for sentence in sentences if 'Foxnews.com'
    updated_sentences = [sentence for sentence in sentences if 'Getty Images
    updated_sentences = [sentence for sentence in sentences if 'CLICK HERE T

    # Join the updated sentences back into content
    updated_content = ' '.join(updated_sentences)

    return updated_content

# Run first iteration of removing the first sentence
df['content'] = df.apply(remove_first_sentence, axis=1)

# Run second iteration to remove  additional noise on some of the rows
df['content'] = df.apply(remove_first_sentence, axis=1)

# Remove prefixes again after the unwanted sentences are removed
df['content'] = df.apply(remove_prefix, axis=1)
```

Let's check to see how our content looks now without the unwanted first and second sentences found in some of the articles.

In [19]:

```
# Checking to see how CNN looks without the unwanted first sentences.
df[df['source']=='cnn']
```

Out[19]:

	source	url	content
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	Chairman of the Joint Chiefs of Staff Gen. CQ ...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Former President Donald Trump has endorsed Nor...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	President Joe Biden and King Abdullah II of Jo...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Former President Donald Trump on Monday asked ...
...	...	...	...
220	cnn	https://www.cnn.com/2024/02/15/politics/navy-f...	Members of Congress pressed the CEO of the nat...
221	cnn	https://www.cnn.com/2024/02/16/politics/kamala...	US Vice President Kamala Harris on Friday call...
222	cnn	https://www.cnn.com/2024/01/30/politics/trump-...	New York state Judge Arthur Engoron has the fu...
223	cnn	https://www.cnn.com/2024/02/15/politics/border...	The acting deputy chief of the US Border Patro...
224	cnn	https://www.cnn.com/2024/02/15/politics/former...	Special counsel David Weiss charged a former F...

143 rows × 3 columns

In [20]:

```
# Checking to see how Fox looks without the unwanted first sentences.
df[df['source']=='foxnews']
```

Out[20]:

	source	url	content
47	foxnews	https://www.foxnews.com/politics/biden-takes-j...	Biden takes jab at special counsel report with...
48	foxnews	https://www.foxnews.com/politics/rfk-jr-apolog...	RFK Jr. drops surprise campaign ad during Supe...
49	foxnews	https://www.foxnews.com/politics/bidens-upcomi...	Biden won't take cognitive test in physical ex...
50	foxnews	https://www.foxnews.com/politics/kamala-harris...	Marc Thiessen questions whether Biden is capab...
51	foxnews	https://www.foxnews.com/politics/climate-activ...	Biden’s export suspension on liquefied natural...
...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republi...	Rep. Ronny Jackson demands Biden take cognitiv...
236	foxnews	https://www.foxnews.com/politics/gop-senators-...	Biden and the Democrats just do not care: Sen....
237	foxnews	https://www.foxnews.com/politics/doj-defends-s...	Former US attorney discusses Special Counsel H...
238	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News’ Politics newsletter with ...
239	foxnews	https://www.foxnews.com/politics/democrats-win...	Dems flipping NY House seat threatens GOP majo...

97 rows × 3 columns

Let's quickly remove references to images embedded into the body content, as it is also noise not needed for topic modeling.

In [21]:

```
# Remove image info

# Define a regular expression pattern to match content inside parentheses
pattern = r'\s*\([^)]*\)'

# Replace content inside parentheses with an empty string
df['content'] = df['content'].str.replace(pattern, '', regex=True)
```

Next, we look at the end of the articles as the content will often end with contributing author information or other material that is not relevant to the topic of the body content. We show the dataframe ending previews and then write a function to remove last sentences if they contain information that is not relevant.

In [22]:
```
# Set pandas to display the full content of a column
pd.set_option('display.max_colwidth', None)
```

In [23]:
```
# Create a new column 'content_end_preview' to show the last part of the cor
df['content_end_preview'] = df['content'].apply(lambda x: x[-500:])
```

In [24]:
```
# Checking the end of CNN articles
df[['source', 'url', 'content_end_preview']] [df['source'] == 'cnn']
```

Out[24]:

	source	url	content_end_preview
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-brown-nato-trump/index.html	eir objectives,” the chairman said. “At the same time, not looking for a broader conflict with the United States.” There have been at least 170 attacks on US and coalition forces in Iraq, Syria and Jordan since October 17. The Pentagon said Monday that those attacks have resulted in 186 wounded or killed in action — including 130 traumatic brain injuries. Three US soldiers were killed in a drone attack in January on a US outpost in Jordan. This story has been updated with additional information.
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-endorse-michael-whatley-lara-trump-rnc/index.html	great job in his home state of North Carolina, and is committed to election integrity, which we must have to keep fraud out of our election so it can’t be stolen,” Trump said in a statement. “My very talented daughter-in-law, Lara Trump, has agreed to run as the RNC Co-Chair. Lara is an extremely talented communicator and is dedicated to all that MAGA stands for. She has told me she wants to accept this challenge and would be GREAT!” he also said. This is a developing story and will be updated.
2	cnn	https://www.cnn.com/2024/02/12/politics/senate-foreign-aid-bill-ukraine/index.html	y be part of the bill, but went on to reject the bipartisan deal amid forceful attacks on the measure by Trump and top House Republicans. Over the weekend, Trump also wrote on Truth Social that the US should stop providing foreign aid unless it is

	source	url	content_end_preview
			structured as a loan, another sign of the political pressure Republicans continue to face amid efforts to send funding to US allies. This story and headline have been updated with additional developments. CNN’s Kate Sullivan contributed to this report.
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens-meeting-with-jordanian-king-comes-at-flashpoint-in-israel-hamas-war/index.html	ions toward an agreement would continue despite the Israeli prime minister’s comments, which Blinken said were referencing the “absolute non-starters” in the proposal. The full Hamas response proposes three phases, each lasting 45 days, including the withdrawal of Israeli troops from Gaza, a massive humanitarian effort, and freedom of movement for people throughout Gaza, according to a copy obtained by CNN. CNN’s MJ Lee, Priscilla Alvarez, Betsy Klein and Kevin Liptak contributed to this report.
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-supreme-court-immunity-filing/index.html	nist ban.” The court may have to decide how it wants to handle the former president’s immunity claim at the same time it is drafting an opinion in the ballot case. Together, the cases have thrust the court into the middle of this year’s presidential election in a way it has largely managed to avoid since its decision in Bush v. Gore effectively decided the 2000 election between former President George W. Bush and former Vice President Al Gore. This

	source	url	content_end_preview
			story has been updated with additional details.
...	...	...	...
220	cnn	https://www.cnn.com/2024/02/15/politics/navy-federal-congressional-black-caucus/index.html	d a separate analysis of public mortgage data by Senate banking committee staff that also found racial disparities in its lending. Navy Federal is also facing a federal class-action lawsuit from mortgage applicants who cite CNN’s reporting and allege that the credit union discriminated against them. A judge approved a motion to consolidate three separate lawsuits against the credit union into a single case last month. Editor’s Note: This story was update to include a statement from Navy Federal.
221	cnn	https://www.cnn.com/2024/02/16/politics/kamala-harris-trump-nato/index.html	NATO allies and abandoned our treaty commitments. Imagine if we went easy on Putin. Let alone encouraged him,” Harris said. “History offers a clue. If we stand by while an aggressor invades its neighbor with impunity, they will keep going. In the case of Putin, that means all of Europe would be threatened. If we fail to impose severe consequences on Russia, other authoritarians across the globe would be emboldened,” Harris said. This story has been updated with additional developments on Friday.
222	cnn	https://www.cnn.com/2024/01/30/politics/trump-fraud-trial-verdict-what-to-watch-for/index.html	oron’s law clerk, leading to the gag order. The judge later extended the order to include Trump’s attorneys from commenting on the judge’s private

	source	url	content_end_preview
			communications with his law clerk. The order does not limit public criticism of Engoron, the district attorney or other parts of the case. As the trial neared an end in December, Engoron reminisced. “In a strange way, I’m gonna miss this trial,” the judge said. “It’s been an experience.” This story has been updated with the anticipated verdict timing.
223	cnn	https://www.cnn.com/2024/02/15/politics/border-patrol-official-suspended-alleged-misconduct/index.html	or off duty. Federal privacy laws prohibit discussion of individual cases.” The Washington Post first reported Martinez’s suspension. Martinez, a 31-year Border Patrol veteran, has also served as the chief patrol agent of the Laredo Sector and deputy chief patrol agent of the Rio Grande Valley Sector, according to the CBP . The US Border Patrol is a law enforcement entity under the umbrella of CBP, in the Department of Homeland Security. CNN’s Piper Hudspeth Blackburn contributed to this report.
224	cnn	https://www.cnn.com/2024/02/15/politics/former-fbi-informant-charged-biden-burisma/index.html	osecutor General was no longer in office,” the indictment states. It continues, “In short, the Defendant transformed his routine and unextraordinary business contacts with Burisma in 2017 and later into bribery allegations against [Joe Biden], the presumptive nominee of one of the two major political parties for President, after expressing bias against [Joe Biden] and his

source	url	content_end_preview
		candidacy.” This story has been updated with additional details. CNN’s Annie Grayer and Manu Raju contributed to this report.

143 rows × 3 columns

In [25]:

```
# Checking the end of Fox News articles
df[['source', 'url', 'content_end_preview']] [df['source'] == 'foxnews']
```

Out[25]:

	source	url	content_end_preview
47	foxnews	https://www.foxnews.com/politics/biden-takes-jab-hur-report-joke-memory-returns-speech-one-more-thing-forgot	ogress." CLICK TO GET THE FOX NEWS APP "The recent Washington Post headline summed it up," Biden added, quoting the newspaper's story titled, "Falling Inflation and Rising Growth Give the U.S. the World's Best Recovery." "The world's best recovery!" Biden said. "It's because you implemented what we did. You made it work." Danielle Wallace is a reporter for Fox News Digital covering politics, crime, police and more. Story tips can be sent to danielle.wallace@fox.com and on Twitter: @danimwallace.
48	foxnews	https://www.foxnews.com/politics/rfk-jr-apologizes-family-super-bowl-ad-claims-no-involvement	icks to stop him. The public sees through it all and won’t stand for it." Kennedy initially sought to challenge President Biden in the 2024 Democratic presidential primary, but the DNC said it would not hold primary debates and stood behind the incumbent president. Fox News' Bradford Betz contributed to this report. Anders Hagstrom is a reporter with Fox News Digital covering national politics and major breaking news events. Send tips to Anders.Hagstrom@Fox.com, or on Twitter: @Hagstrom_Anders.
49	foxnews	https://www.foxnews.com/politics/bidens-upcoming-physical-exam-will-not-include-cognitive-test-white-house-says	has been my experience with this president," she said. Biden's age is a major concern among U.S. voters, 86% of whom say he is too old to serve a second term, according to an ABC poll. A Sunday poll from ABC/Ipsos found that 86% of Americans believe Biden is too old to serve another term, including 73% of Democrats. Anders Hagstrom is a reporter with Fox News Digital covering

	source	url	content_end_preview
			national politics and major breaking news events. Send tips to Anders.Hagstrom@Fox.com, or on Twitter: @Hagstrom_Anders.
50	foxnews	https://www.foxnews.com/politics/kamala-harris-ready-serve-democrats-sound-alarm-about-bidens-age	s crying and wet the bed," Begala quipped on CNN last Friday. "This is terrible for Democrats. And anybody with a functioning brain knows that," he declared. GOP CAMPAIGN ARM LAUNCHES MEDIA BLITZ AGAINST DEMS WHO OPPOSED VIOLENT CRIME BILL AS CRISIS IN DC SPIRALS Then-Democrat presidential candidate Hillary Clinton makes a concession speech after being defeated by Donald Trump in New York on November 9, 2016. Brandon Gillespie is an associate editor at Fox News. Follow him on X at @BGillespieAL.
51	foxnews	https://www.foxnews.com/politics/climate-activists-arrested-shutting-down-biden-campaign-hq	power plant electricity generation, push electric vehicles and incentivize the electrification of the residential sector. "I mean, it literally is the existential threat. It’s even more consequential than nuclear power, nuclear war," he added. "That would be horrible and awful, and it would just make the environment incredibly worse. But it’s about the environment." The Biden campaign didn't immediately respond to a request for comment. Thomas Catenacci is a politics writer for Fox News Digital.
...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republicans-push-biden-take-cognitive-test-hur-report-obvious-mental-decline	. Nick" when reached by Fox News Digital on Friday morning. Earlier this week, White House press secretary Karine Jean-Pierre

	source	url	content_end_preview
			told reporters that Biden would not be taking a cognitive test as part of his regular physical exam. Elizabeth Elkind is a reporter for Fox News Digital focused on Congress as well as the intersection of Artificial Intelligence and politics. Previous digital bylines seen at Daily Mail and CBS News. Follow on Twitter at @liz_elkind and send tips to elizabeth.elkind@fox.com
236	foxnews	https://www.foxnews.com/politics/gop-senators-urge-biden-admin-to-end-racial-discrimination-policy-in-chips-grants-before-it-breaks-the-law	es to discriminate on the basis of race when making and enforcing contracts." Sen. Cynthia Lummis, R-Wyo., conducts a news conference Aug. 9. 29. If she fails to rescind the policy, Cruz and his colleagues are demanding that she detail "the reasons you believe the Guidance does not violate the United States Constitution or Title VI, or induce private parties to violate Section 1981." Brooke Singman is a political correspondent and reporter for Fox News Digital, Fox News Channel and FOX Business.
237	foxnews	https://www.foxnews.com/politics/doj-defends-special-counsel-report-bidens-memory-consistent-legal-requirement-not-gratuitous	later revealed to Fox News that it was Biden who brought up his son's 2015 death-not Hur. "We conclude that no criminal charges are warranted in this matter," the report, released Thursday, states. Fox News' David Spunt and Jake Gibson contributed to this report. Sarah Rumpf-Whitten is a breaking news writer for Fox News Digital and Fox Business. She is a native of Massachusetts and is based in Orlando, Florida. Story tips and ideas can be sent to sarah.rumpf@fox.com and on X: @s_rumpfwhitten .

	source	url	content_end_preview
238	foxnews	https://www.foxnews.com/politics/fox-news-politics-trump-vows-appeal	run for president, serve as Manchin's VP ...Read more 'COMMONSENSE CONSERVATIVE': Former special forces soldier lands big endorsement in race to flip House seat ...Read more 'RACE OF HIS LIFE' : Dem Sen blasts GOP for not caring about immigration; record comes back to haunt him ...Read more Subscribe now to get Fox News Politics newsletter in your inbox. Get the latest updates from the 2024 campaign trail, exclusive interviews and more on FoxNews.com . This article was written by Fox News staff.
239	foxnews	https://www.foxnews.com/politics/democrats-win-seat-republicans-win-impeachment-two-presidents-clash-over-nato	airing his personal criticism of the president’s mental acuity. And everyone is getting sustained exposure to a system that generally favors political maneuvering over actual results. Get the latest updates from the 2024 campaign trail, exclusive interviews and more at our Fox News Digital election hub. Howard Kurtz is the host of FOX News Channel's MediaBuzz. Based in Washington, D.C., he joined the network in July 2013 and regularly appears on Special Report with Bret Baier and other programs.

97 rows × 3 columns

We see that indeed some articles end with information about the authors or otherwise irrelevant information. Below is our function to handle some of the instances. We need to run it multiple times, as each time it is run there is a new last sentence that counts as noise that we want to get rid of in some instances.

In [26]:

```
def remove_last_sentence(row):
    sentences = row['content'].split('. ')

    if len(sentences) > 1:  # Check if there's more than one sentence
        last_sentence = sentences[-1]  # Get the last sentence
        if ('This story has been updated with additional information.' in la
            'contributed to this' in last_sentence or
            'will be updated' in last_sentence or
            'have been updated' in last_sentence or
            'APP Fox News' in last_sentence or
            'Fox News' in last_sentence or
            'FoxNews.com' in last_sentence or
            '@Fox.com' in last_sentence or
            'Fox News Digital' in last_sentence or
            'Fox News Channel and FOX Business' in last_sentence or

            'This story has been updated with additional reaction' in last_s
            'This report has been updated with additional information' in la
            'who covers politics' in last_sentence or
            'follow him on' in last_sentence or
            'Follow him on' in last_sentence or
            '@fox.com' in last_sentence or
            '@Fox.com' in last_sentence or
            'FoxNews.com' in last_sentence or
            'Fox News Digital' in last_sentence or
            'contributed to this' in last_sentence or
            'Politics newsletter' in last_sentence or
            'Fox News Digital' in last_sentence or
            'email' in last_sentence):

            updated_content = '. '.join(sentences[:-1])  # Join all sentence
            return updated_content

    return row['content']

# Apply the function to the DataFrame four times
df['content'] = df.apply(remove_last_sentence, axis=1)
df['content'] = df.apply(remove_last_sentence, axis=1)
df['content'] = df.apply(remove_last_sentence, axis=1)
df['content'] = df.apply(remove_last_sentence, axis=1)
df['content_end_preview'] = df['content'].apply(lambda x: x[-500:])
```

In [27]:

```
# Checking the end of Fox News articles after we run our function
df[['source', 'url', 'content_end_preview']] [df['source'] == 'foxnews']
```

Out[27]:

	source	url	content_end_preview
47	foxnews	https://www.foxnews.com/politics/biden-takes-jab-hur-report-joke-memory-returns-speech-one-more-thing-forgot	said, referring to companies charging the same amount for a product while reducing quantity. "I'm calling on corporations to pass their savings on to consumers, for God's sake. We're making real progress." CLICK TO GET THE FOX NEWS APP "The recent Washington Post headline summed it up," Biden added, quoting the newspaper's story titled, "Falling Inflation and Rising Growth Give the U.S. the World's Best Recovery." "The world's best recovery!" Biden said. "It's because you implemented what we did
48	foxnews	https://www.foxnews.com/politics/rfk-jr-apologizes-family-super-bowl-ad-claims-no-involvement	ever wars, and chronic disease. RFK Jr. offers us real change along with freedom, trust and hope. Like his uncle and his father, Kennedy is a corruption fighter, and it's no wonder the DNC is trying every old trick and inventing new tricks to stop him. The public sees through it all and won’t stand for it." Kennedy initially sought to challenge President Biden in the 2024 Democratic presidential primary, but the DNC said it would not hold primary debates and stood behind the incumbent president
49	foxnews	https://www.foxnews.com/politics/bidens-upcoming-physical-exam-will-not-include-cognitive-test-white-house-says	and continues to find him to be "sharp" and "on top of things." "When we have meetings with him and his staff he is constantly pushing us, trying to get more information, and so that has been my experience with this president," she said. Biden's age is a major concern among U.S. voters, 86% of whom say he is too old to serve a second term, according to an ABC poll. A

	source	url	content_end_preview
			Sunday poll from ABC/Ipsos found that 86% of Americans believe Biden is too old to serve another term, including 73% of Democrats
50	foxnews	https://www.foxnews.com/politics/kamala-harris-ready-serve-democrats-sound-alarm-about-bidens-age	Look, I'm a Biden supporter, and I slept like a baby last night: I woke up every two hours crying and wet the bed," Begala quipped on CNN last Friday. "This is terrible for Democrats. And anybody with a functioning brain knows that," he declared. GOP CAMPAIGN ARM LAUNCHES MEDIA BLITZ AGAINST DEMS WHO OPPOSED VIOLENT CRIME BILL AS CRISIS IN DC SPIRALS Then-Democrat presidential candidate Hillary Clinton makes a concession speech after being defeated by Donald Trump in New York on November 9, 2016
51	foxnews	https://www.foxnews.com/politics/climate-activists-arrested-shutting-down-biden-campaign-hq	n onslaught of environmental regulations to curb fossil fuel power plant electricity generation, push electric vehicles and incentivize the electrification of the residential sector. "I mean, it literally is the existential threat. It's even more consequential than nuclear power, nuclear war," he added. "That would be horrible and awful, and it would just make the environment incredibly worse. But it's about the environment." The Biden campaign didn't immediately respond to a request for comment
...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republicans-push-biden-take-cognitive-test-hur-report-obvious-mental-decline	al practices and maiming patients. The White House sent another image of "Dr. Nick" when reached by Fox

	source	url	content_end_preview
			News Digital on Friday morning. Earlier this week, White House press secretary Karine Jean-Pierre told reporters that Biden would not be taking a cognitive test as part of his regular physical exam. Elizabeth Elkind is a reporter for Fox News Digital focused on Congress as well as the intersection of Artificial Intelligence and politics. Previous digital bylines seen at Daily Mail and CBS News
236	foxnews	https://www.foxnews.com/politics/gop-senators-urge-biden-admin-to-end-racial-discrimination-policy-in-chips-grants-before-it-breaks-the-law	race of their suppliers. Title VI forbids such discrimination," they wrote. "In addition to instructing the federal government to violate the law, the Guidance also encourages private businesses to discriminate on the basis of race in violation of federal law, specifically Section 1981," they continued. "Section 1981 makes it illegal for private companies to discriminate on the basis of race when making and enforcing contracts." Sen. Cynthia Lummis, R-Wyo., conducts a news conference Aug. 9. 29
237	foxnews	https://www.foxnews.com/politics/doj-defends-special-counsel-report-bidens-memory-consistent-legal-requirement-not-gratuitous	tten notes while eating breakfast alongside senators on January 29, 2015. Two sources later revealed to Fox News that it was Biden who brought up his son's 2015 death-not Hur. "We conclude that no criminal charges are warranted in this matter," the report, released Thursday, states. Fox News' David Spunt and Jake Gibson contributed to this report. Sarah Rumpf-Whitten is a breaking news writer for Fox News Digital and Fox Business. She is a

	source	url	content_end_preview
			native of Massachusetts and is based in Orlando, Florida
238	foxnews	https://www.foxnews.com/politics/fox-news-politics-trump-vows-appeal	he former president inflated his assets and committed fraud. Trump spoke exclusively to Fox News Digital shortly after Engoron's ruling was made public Friday afternoon. "A crooked New York judge working with the very corrupt attorney general of New York State, who ran on the basis of ‘I will get trump’ before knowing me — before even knowing anything about me — just ruled that I have to pay a fine of $355 million based on absolutely nothing," Trump told Fox News Digital. "No victims. No damages
239	foxnews	https://www.foxnews.com/politics/democrats-win-seat-republicans-win-impeachment-two-presidents-clash-over-nato	airing his personal criticism of the president’s mental acuity. And everyone is getting sustained exposure to a system that generally favors political maneuvering over actual results. Get the latest updates from the 2024 campaign trail, exclusive interviews and more at our Fox News Digital election hub. Howard Kurtz is the host of FOX News Channel's MediaBuzz. Based in Washington, D.C., he joined the network in July 2013 and regularly appears on Special Report with Bret Baier and other programs.

97 rows × 3 columns

In [28]:
```
# Checking the end of CNN articles after we run our function
df[['source', 'url', 'content_end_preview']] [df['source'] == 'cnn']
```

Out[28]:

	source	url	content_end_preview
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-brown-nato-trump/index.html	r militia groups and others to put pressure, to achieve their objectives," the chairman said. "At the same time, not looking for a broader conflict with the United States." There have been at least 170 attacks on US and coalition forces in Iraq, Syria and Jordan since October 17. The Pentagon said Monday that those attacks have resulted in 186 wounded or killed in action — including 130 traumatic brain injuries. Three US soldiers were killed in a drone attack in January on a US outpost in Jordan
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-endorse-michael-whatley-lara-trump-rnc/index.html	has been with me from the beginning, has done a great job in his home state of North Carolina, and is committed to election integrity, which we must have to keep fraud out of our election so it can't be stolen," Trump said in a statement. "My very talented daughter-in-law, Lara Trump, has agreed to run as the RNC Co-Chair. Lara is an extremely talented communicator and is dedicated to all that MAGA stands for. She has told me she wants to accept this challenge and would be GREAT!" he also said
2	cnn	https://www.cnn.com/2024/02/12/politics/senate-foreign-aid-bill-ukraine/index.html	would have combined the foreign aid with a bipartisan border deal. Republicans had initially demanded that border security be part of the bill, but went on to reject the bipartisan deal amid forceful attacks on the measure by Trump and top House Republicans.

	source	url	content_end_preview
			Over the weekend, Trump also wrote on Truth Social that the US should stop providing foreign aid unless it is structured as a loan, another sign of the political pressure Republicans continue to face amid efforts to send funding to US allies
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens-meeting-with-jordanian-king-comes-at-flashpoint-in-israel-hamas-war/index.html	stage deal in Gaza “delusional.” Secretary of State Antony Blinken previously said negotiations toward an agreement would continue despite the Israeli prime minister’s comments, which Blinken said were referencing the “absolute non-starters” in the proposal. The full Hamas response proposes three phases, each lasting 45 days, including the withdrawal of Israeli troops from Gaza, a massive humanitarian effort, and freedom of movement for people throughout Gaza, according to a copy obtained by CNN
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-supreme-court-immunity-filing/index.html	nist ban.” The court may have to decide how it wants to handle the former president’s immunity claim at the same time it is drafting an opinion in the ballot case. Together, the cases have thrust the court into the middle of this year’s presidential election in a way it has largely managed to avoid since its decision in Bush v. Gore effectively decided the 2000 election between former President George W. Bush and former Vice President Al Gore. This story has been updated with additional details.

	source	url	content_end_preview
...	...	...	...
220	cnn	https://www.cnn.com/2024/02/15/politics/navy-federal-congressional-black-caucus/index.html	d a separate analysis of public mortgage data by Senate banking committee staff that also found racial disparities in its lending. Navy Federal is also facing a federal class-action lawsuit from mortgage applicants who cite CNN’s reporting and allege that the credit union discriminated against them. A judge approved a motion to consolidate three separate lawsuits against the credit union into a single case last month. Editor’s Note: This story was update to include a statement from Navy Federal.
221	cnn	https://www.cnn.com/2024/02/16/politics/kamala-harris-trump-nato/index.html	NATO allies and abandoned our treaty commitments. Imagine if we went easy on Putin. Let alone encouraged him,” Harris said. “History offers a clue. If we stand by while an aggressor invades its neighbor with impunity, they will keep going. In the case of Putin, that means all of Europe would be threatened. If we fail to impose severe consequences on Russia, other authoritarians across the globe would be emboldened,” Harris said. This story has been updated with additional developments on Friday.
222	cnn	https://www.cnn.com/2024/01/30/politics/trump-fraud-trial-verdict-what-to-watch-for/index.html	oron’s law clerk, leading to the gag order. The judge later extended the order to include Trump’s attorneys from commenting on the judge’s private communications with his law clerk. The order does not limit public criticism

	source	url	content_end_preview
			of Engoron, the district attorney or other parts of the case. As the trial neared an end in December, Engoron reminisced. “In a strange way, I’m gonna miss this trial,” the judge said. “It’s been an experience.” This story has been updated with the anticipated verdict timing.
223	cnn	https://www.cnn.com/2024/02/15/politics/border-patrol-official-suspended-alleged-misconduct/index.html	“This is the case whether the alleged misconduct occurs on or off duty. Federal privacy laws prohibit discussion of individual cases.” The Washington Post first reported Martinez’s suspension. Martinez, a 31-year Border Patrol veteran, has also served as the chief patrol agent of the Laredo Sector and deputy chief patrol agent of the Rio Grande Valley Sector, according to the CBP . The US Border Patrol is a law enforcement entity under the umbrella of CBP, in the Department of Homeland Security
224	cnn	https://www.cnn.com/2024/02/15/politics/former-fbi-informant-charged-biden-burisma/index.html	Biden] had no ability to influence U.S. policy and when the Prosecutor General was no longer in office,” the indictment states. It continues, “In short, the Defendant transformed his routine and unextraordinary business contacts with Burisma in 2017 and later into bribery allegations against [Joe Biden], the presumptive nominee of one of the two major political parties for President, after expressing bias against [Joe Biden] and his candidacy.” This story

source	url	content_end_preview
		has been updated with additional details

143 rows × 3 columns

Now that we have finished setting up our remove_last_sentence function, we can remove the "content_end_preview" column and reset the pandas display setting.

In [29]:
```
# Drop the content end preview column
df.drop(columns=['content_end_preview'], inplace=True)

# Reset the columns so that we go back to truncating the "content" column
pd.reset_option('display.max_colwidth')
```

Standardizing Entity Names

Let's start by combining specified word pairs so that we handle cases where two or more words refer to a single entity, such as "Hunter Biden" or "Supreme Court."

In [30]:
```
# Combine specified word pairs

df['content'] = df['content'].str.replace(r'\bHunter\s+Biden\b', 'HunterBide
df['content'] = df['content'].str.replace(r'\bHUNTER\s+Biden\b', 'HunterBide
df['content'] = df['content'].str.replace(r'\bSouth\s+Carolina\b', 'SouthCar
df['content'] = df['content'].str.replace(r'\bSupreme\s+Court\b', 'SupremeCo
df['content'] = df['content'].str.replace(r'\bsupreme\s+court\b', 'SupremeCo
df['content'] = df['content'].str.replace(r'\bCourt\s+House\b', 'CourtHouse'
df['content'] = df['content'].str.replace(r'\bcourt\s+house\b', 'CourtHouse'
df['content'] = df['content'].str.replace(r'\bHouse\s+Representative\b', 'Ho
df['content'] = df['content'].str.replace(r'\bhouse\s+representative\b', 'Ho
df['content'] = df['content'].str.replace(r'\bHouse\s+Rep\b', 'HouseRep', re
df['content'] = df['content'].str.replace(r'\bhouse\s+rep\b', 'HouseRep', re
df['content'] = df['content'].str.replace(r'\bvoters\b', 'voter', regex=True
df['content'] = df['content'].str.replace(r'\bvotes\b', 'vote', regex=True)
df['content'] = df['content'].str.replace(r'\bdemocratic(?:s)?\b', 'Democrat
df['content'] = df['content'].str.replace(r'\bDemocrats\b', 'Democrat', rege
df['content'] = df['content'].str.replace(r'\brepublicans\b', 'Republican', 
df['content'] = df['content'].str.replace(r'\bRepublicans\b', 'Republican', 
df['content'] = df['content'].str.replace(r'\bwhite\s+house\b', 'WhiteHouse'
df['content'] = df['content'].str.replace(r'\bWhite\s+house\b', 'WhiteHouse'
df['content'] = df['content'].str.replace(r'\bNew\s+York\b', 'NewYork', rege
```

We now need to account for variations in Biden's and Trump's names. This is because we want the model to see the different spellings as referring to the same thing. We accomplish this with the following code.

In [31]:
```
biden_variations = df['content'].str.findall(
    r'\bPresident\s+Joe\s+Biden\b|'
    r'\bPresident\s+Biden\b|'
    r'\bJoe\s+Biden(?:’s)?\b|'
    r'\bBiden(?:’s|s)?\b|'
    r'\bBIDEN\b|'
    r'\bBiden\'s\b'
)
# Flatten the list of variations
biden_variations = [item for sublist in biden_variations for item in sublist

# Count occurrences of each variation
biden_variation_counts = Counter(biden_variations)

# Replace variations of Biden's name with 'Biden' in the content column
df['content'] = df['content'].str.replace(
    r'\bPresident\s+Joe\s+Biden\b|'
    r'\bPresident\s+Biden\b|'
    r'\bJoe\s+Biden(?:’s)?\b|'
    r'\bBiden(?:’s|s)?\b|'
    r'\bBIDEN\b|'
    r'\bBiden\'s\b'
    , 'Biden', regex=True)

print("Occurrences of different variations of Biden's name:")
for variation, count in biden_variation_counts.items():
    print(f"{variation}: {count}")
```

```
Occurrences of different variations of Biden's name:
President Joe Biden: 116
Biden: 861
Biden’s: 156
President Biden: 159
Bidens: 23
Joe Biden: 137
Joe Biden’s: 13
President Joe Biden: 1
BIDEN: 105
```

In [32]:

```
# Count occurrences of 'Biden' after replacement
biden_count_after = df['content'].str.count('Biden').sum()

print("Occurrences of Biden after replacement:", biden_count_after)
```

```
Occurrences of Biden after replacement: 1709
```

In [33]:

```
# Find all variations of Trump's name in the content column
trump_variations = df['content'].str.findall(
    r'\bPresident\s+Donald\s+Trump\b|'
    r'\bPresident\s+Trump\b|'
    r'\bDonald\s+Trump(?:’s)?\b|'
    r'\bTrump(?:’s)?\b|'
    r'\bTRUMP(?:’S)?\b|'
    r'\bFormer\s+President\s+Donald\s+Trump\b|'
    r'\bDonald\s+J(?:ohn)?\s+Trump\b'
)

# Flatten the list of variations
trump_variations = [item for sublist in trump_variations for item in sublist

# Count occurrences of each variation
```

```
trump_variation_counts = Counter(trump_variations)

# Replace variations of Trump's name with 'Trump' in the content column
df['content'] = df['content'].str.replace(
    r'\bPresident\s+Donald\s+Trump\b|'
    r'\bPresident\s+Trump\b|'
    r'\bDonald\s+Trump(?:’s)?\b|'
    r'\bTrump(?:’s)?\b|'
    r'\bTRUMP(?:’S)?\b|'
    r'\bFormer\s+President\s+Donald\s+Trump\b|'
    r'\bDonald\s+J(?:ohn)?\s+Trump\b'
    , 'Trump', regex=True)

print("Occurrences of different variations of Trump's name:")
for variation, count in trump_variation_counts.items():
    print(f"{variation}: {count}")
```

```
Occurrences of different variations of Trump's name:
President Donald Trump: 78
Trump’s: 380
Trump: 1259
Former President Donald Trump: 32
President Trump: 81
Donald Trump: 90
Donald Trump’s: 15
Donald Trump: 1
TRUMP: 35
TRUMP’S: 2
President Trump: 1
```

In [34]:
```
# Count occurrences of 'Trump' after replacement
trump_count_after = df['content'].str.count('Trump').sum()

print("Occurrences of Trump after replacement:", trump_count_after)
```

```
Occurrences of Trump after replacement: 1977
```

Data Preprocessing Pipeline

Next, we remove stop words and punctuation, and then we tokenize and prepare data for use in the model.

In [35]:
```
punctuation = set(punctuation) # speeds up comparison
sw = stopwords.words("english")
extra_sw = ['cnn', 'fox', 'news', 'said', '—', '–', '——', '—','told', 'would
            'also', "it’s", 'think', 'time', 'even', 'former', 'party', 'i',
            'images', 'getty', 'im', 'this', 'we', 'it', 'digital', 'the', '
sw.extend(extra_sw)
whitespace_pattern = re.compile(r"\s+")

def remove_stop(tokens) :

    return [t for t in tokens if t.lower() not in sw]

def remove_punctuation(text, punct_set=punctuation) :
    return("".join([ch for ch in text if ch not in punct_set]))

def tokenize(text) :

    return re.split(whitespace_pattern, text)

def prepare(text, pipeline) :

    tokens = str(text)

    for transform in pipeline :
        tokens = transform(tokens)

    return(tokens)

pipeline = [str.lower, remove_punctuation, tokenize, remove_stop]
```

In [36]:
```
# Tokenize and preprocess each row
df['tokens'] = df['content'].apply(lambda x: prepare(x, pipeline=pipeline))

# Print the resulting dataframe
df.head()
```

Out[36]:

	source	url	content	tokens
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	Chairman of the Joint Chiefs of Staff Gen. CQ ...	[chairman, joint, chiefs, staff, gen, cq, brow...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump has endorsed North Carolina Republican P...	[trump, endorsed, north, carolina, republican,...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...	[senate, inching, closer, final, passage, 953,...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Biden and King Abdullah II of Jordan met Monda...	[biden, king, abdullah, ii, jordan, met, monda...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump on Monday asked the SupremeCourt to step...	[trump, monday, asked, supremecourt, step, cha...

Basic Descriptive Statistics

Below we write a function that allows us to view the results of our preprocessed data from CNN and Fox News. We find the total words (tokens), unique words, total characters, lexical diversity, and most common words for each news organization.

In [37]:
```
def descriptive_stats(tokens, num_tokens = 50, verbose=True) :
    """
        Given a list of tokens, print number of tokens, number of unique tok
        number of characters, lexical diversity (https://en.wikipedia.org/wi
        and num_tokens most common tokens. Return a list with the number of
        of unique tokens, lexical diversity, and number of characters.

    """
    num_tokens = len(tokens)
    num_unique_tokens = len(set(tokens))
    lexical_diversity = num_unique_tokens / num_tokens
    num_characters = sum(len(s) for s in tokens)

    if verbose :
        print(f"There are {num_tokens} tokens in the data.")
        print(f"There are {num_unique_tokens} unique tokens in the data.")
        print(f"There are {num_characters} characters in the data.")
        print(f"The lexical diversity is {lexical_diversity:.3f} in the data
        print (f"The ten most common words are:")
        print(Counter(tokens).most_common(10))

    return([num_tokens, num_unique_tokens,
            lexical_diversity,
            num_characters])
```

In [38]:
```
# calls to descriptive_stats here

print("CNN News Stats\n")

descriptive_stats(
    [token for tokens in df[df['source'] == 'cnn']['tokens']for token in tok

print('\n')
print("FoxNews Stats\n")

descriptive_stats(
    [token for tokens in df[df['source'] == 'foxnews']['tokens']for token ir
```

```
CNN News Stats

There are 76965 tokens in the data.
There are 11340 unique tokens in the data.
There are 507542 characters in the data.
The lexical diversity is 0.147 in the data.
The ten most common words are:
[('trump', 1527), ('biden', 741), ('republican', 560), ('house', 459), ('u
s', 411), ('president', 402), ('election', 396), ('democrat', 360), ('case',
284), ('campaign', 254)]


FoxNews Stats

There are 40271 tokens in the data.
There are 7106 unique tokens in the data.
There are 262325 characters in the data.
The lexical diversity is 0.176 in the data.
The ten most common words are:
[('biden', 679), ('house', 403), ('trump', 330), ('republican', 245), ('pres
ident', 224), ('democrat', 193), ('us', 187), ('senate', 167), ('security',
155), ('special', 153)]
```

```
Out[38]: [40271, 7106, 0.17645452062278066, 262325]
```

Saving the Data

Below we create a csv file to use for modeling.

```
In [39]: # save df for next step

         df.to_csv('data/cleaned.csv', index=False)
```

MSADS509 Final Project Modeling

In [71]:

```
import datetime
import random
import time
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin
from collections import defaultdict, Counter

import pandas as pd
import numpy as np
from tqdm.auto import tqdm
import os
import re
import spacy
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud

import pyLDAvis
import pyLDAvis.lda_model
import pyLDAvis.gensim_models

from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA, NMF, TruncatedSVD, LatentDirichletAllocation
from sklearn.metrics import accuracy_score, classification_report

from scipy.sparse import hstack

from spacy.lang.en.stop_words import STOP_WORDS as stopwords

from pandas import json_normalize

nlp = spacy.load('en_core_web_sm')

from nltk.corpus import stopwords
from string import punctuation
from nltk.tokenize import word_tokenize
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.stem import PorterStemmer
import nltk
nltk.download('punkt')
nltk.download('vader_lexicon')

import warnings
warnings.filterwarnings("ignore")
```

```
[nltk_data] Downloading package punkt to /Users/kevinbaum/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/kevinbaum/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

Load data from Data folder

In [72]:

```
# make sure your directory is the same one that was used to store the cleaned dataframe
df = pd.read_csv('Data/cleaned.csv')

def clean_tokens(tokens):
    return [token.strip("[]'") for token in tokens.split(', ')]

df['tokens'] = df['tokens'].apply(clean_tokens)
df
```

Out[72]:

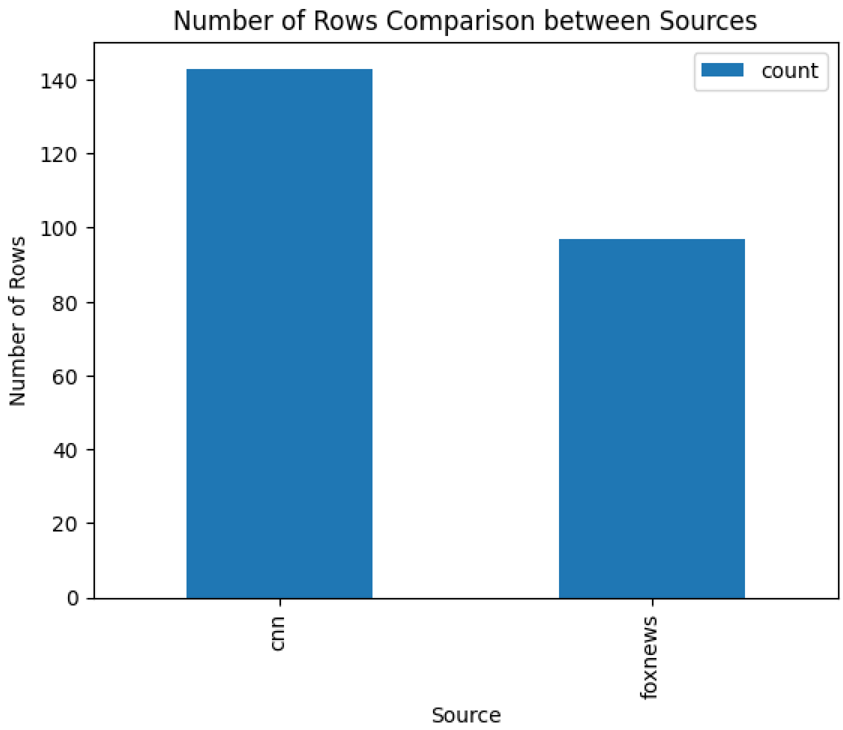
	source	url	content	tokens
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	Chairman of the Joint Chiefs of Staff Gen. CQ ...	[chairman, joint, chiefs, staff, gen, cq, brow...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump has endorsed North Carolina Republican P...	[trump, endorsed, north, carolina, republican,...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...	[senate, inching, closer, final, passage, 953,...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Biden and King Abdullah II of Jordan met Monda...	[biden, king, abdullah, ii, jordan, met, monda...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump on Monday asked the SupremeCourt to step...	[trump, monday, asked, supremecourt, step, cha...
...	...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republi...	Rep. Ronny Jackson demands Biden take cognitiv...	[rep, ronny, jackson, demands, biden, take, co...
236	foxnews	https://www.foxnews.com/politics/gop-senators-...	Biden and the Democrat just do not care: Sen. ...	[biden, democrat, care, sen, ted, cruz, sen, t...
237	foxnews	https://www.foxnews.com/politics/doj-defends-s...	Former US attorney discusses Special Counsel H...	[us, attorney, discusses, special, counsel, hu...
238	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News' Politics newsletter with ...	[welcome, news', politics, newsletter, latest,...
239	foxnews	https://www.foxnews.com/politics/democrats-win...	Dems flipping NY House seat threatens GOP majo...	[dems, flipping, ny, house, seat, threatens, g...

240 rows × 4 columns

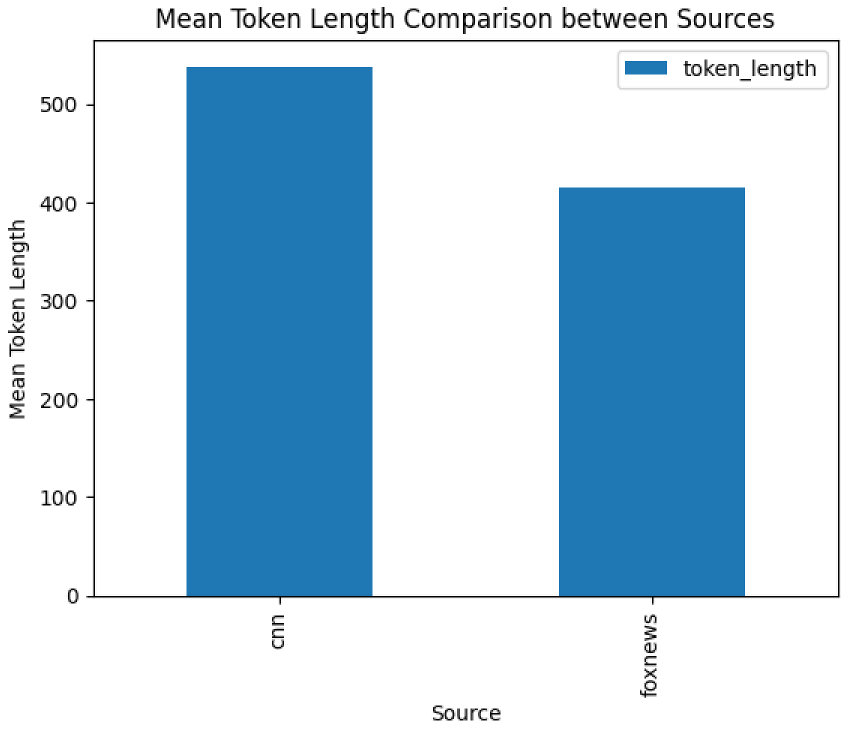
EDA for tokens

In [73]:

```
# Count the number of rows for each source
source_counts = df['source'].value_counts()

source_counts.plot(kind='bar', legend=True)
plt.xlabel('Source')
plt.ylabel('Number of Rows')
plt.title('Number of Rows Comparison between Sources')
plt.show()
```

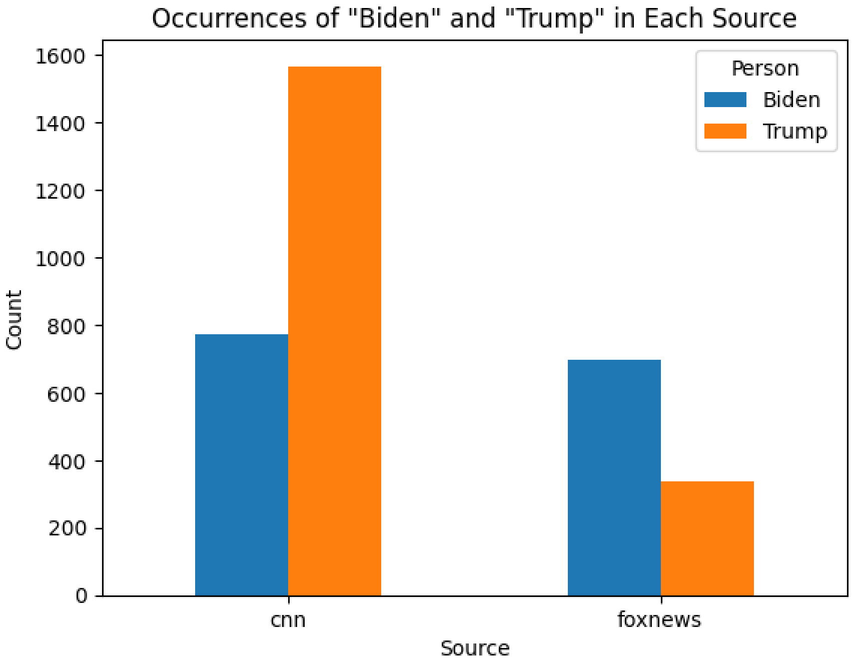


In [74]:
```python
# the length of tokens for each article
df['token_length'] = df['tokens'].apply(lambda x: len(x))

source_token_length = df.groupby('source')['token_length'].mean()

source_token_length.plot(kind='bar', legend=True)
plt.xlabel('Source')
plt.ylabel('Mean Token Length')
plt.title('Mean Token Length Comparison between Sources')
plt.show()
```



In [75]:
```python
# count occurrences of a word in a list
def count_occurrences(tokens, word):
    return sum(1 for token in tokens if re.search(r'\b{}\b'.format(word), token, flags=re.IGNORECASE))
```

```
biden_counts = df.groupby('source')['tokens'].apply(lambda x: sum(count_occurrences(tokens, 'biden') for t
trump_counts = df.groupby('source')['tokens'].apply(lambda x: sum(count_occurrences(tokens, 'trump') for t

counts_df = pd.DataFrame({'Biden': biden_counts, 'Trump': trump_counts})

counts_df.plot(kind='bar')
plt.xlabel('Source')
plt.ylabel('Count')
plt.title('Occurrences of "Biden" and "Trump" in Each Source')
plt.xticks(rotation=0)
plt.legend(title='Person')
plt.show()
```

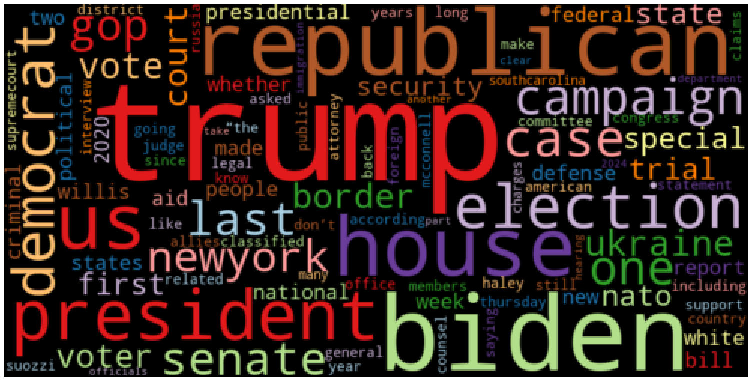


WordCloud for each Source

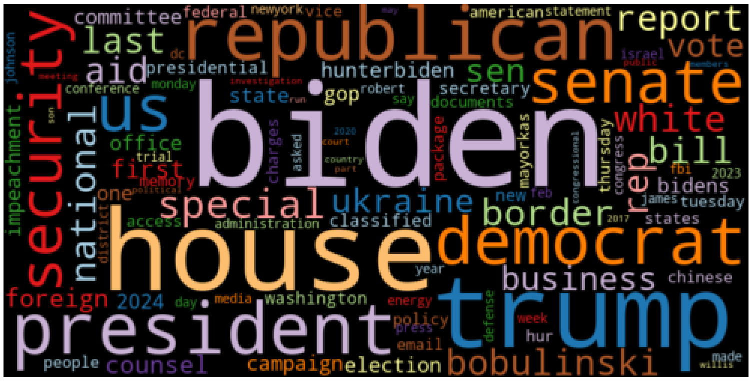
In [76]:

```
def wordcloud(df, title=None, max_words=100, stopwords=None):
    unique_sources = df['source'].unique()
    for source in unique_sources:
        tokens = df[df['source'] == source]['tokens']
        all_tokens = [token for sublist in tokens for token in sublist]
        counter = Counter(all_tokens)

        # Filter stop words in frequency counter
        if stopwords is not None:
            counter = {token: freq for (token, freq) in counter.items() if token not in stopwords}

        wc = WordCloud(width=800, height=400,
                       background_color="black", colormap="Paired",
                       max_font_size=150, max_words=max_words)
        wc.generate_from_frequencies(counter)

        plt.title(f"{title} - {source}")
        plt.imshow(wc, interpolation='bilinear')
        plt.axis("off")
        plt.show()

wordcloud(df, title="Top 100 Popular Words")
```

Top 100 Popular Words - cnn

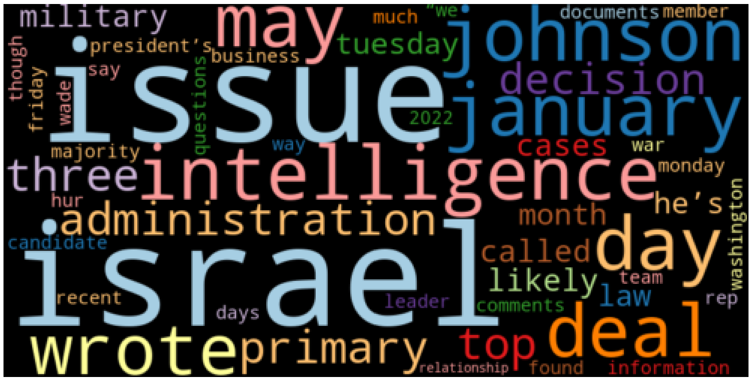


Top 100 Popular Words - foxnews

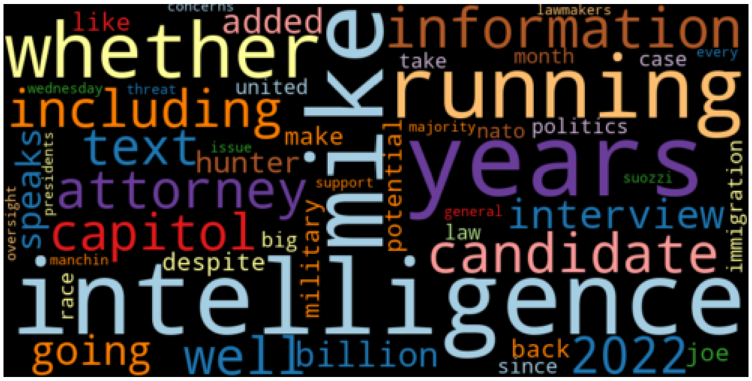


In [77]:
```
def wordcloud(df, title=None, max_words=200, stopwords=None):
    unique_sources = df['source'].unique()
    for source in unique_sources:
        tokens = df[df['source'] == source]['tokens']
        all_tokens = [token for sublist in tokens for token in sublist]
        counter = Counter(all_tokens)

        # Filter stop words in frequency counter
        if stopwords is not None:
            counter = {token: freq for (token, freq) in counter.items() if token not in stopwords}

        # Sort the counter by frequency and get the words ranked from 101st to 150th
        sorted_counter = dict(counter.most_common())
        words_101_to_150 = dict(list(sorted_counter.items())[100:150])

        wc = WordCloud(width=800, height=400,
                       background_color="black", colormap="Paired",
                       max_font_size=150, max_words=max_words)
        wc.generate_from_frequencies(words_101_to_150)

        plt.title(f"{title} - {source}")
        plt.imshow(wc, interpolation='bilinear')
        plt.axis("off")
        plt.show()

wordcloud(df, title="Words Ranked 101-150 by Occurrences")
```

Words Ranked 101-150 by Occurrences - cnn



Words Ranked 101-150 by Occurrences - foxnews



Modeling and Model Evaluations

Topic Modeling

```
In [78]: sw = stopwords.words("english")
punctuation = set(punctuation) # speeds up comparison
extra_sw = ['cnn', 'fox', 'news', 'said', '-', '–', '––', '—','told', 'would', 'read', 'get', 'could',
            'also', "it's", 'think', 'time', 'even', 'former', 'party', 'i', '“i', "she's", 'says',
            'images', 'getty', 'im', 'this', 'we', 'it', 'digital', 'the', 'that', 'story', 'doesn']
sw.extend(extra_sw)
```

```
In [79]: # define the topic displaying function
def display_topics(model, features, no_top_words=5):
    for topic, words in enumerate(model.components_):
        total = words.sum()
        largest = words.argsort()[::-1] # invert sort order
        print("\nTopic %02d" % topic)
        for i in range(0, no_top_words):
            print("  %s (%2.2f)" % (features[largest[i]], abs(words[largest[i]]*100.0/total)))
```

```
In [80]: # create separate dataframes for the two news sources
cnn_df = df[df['source'] == 'cnn']
fox_df = df[df['source'] == 'foxnews']

# define the function to join tokens back into a string
def join_tokens(tokens):
    return ' '.join(tokens)

# Apply the join_tokens function to the "tokens" column
cnn_df["tokens_str"] = cnn_df["tokens"].apply(join_tokens)
fox_df["tokens_str"] = fox_df["tokens"].apply(join_tokens)
```

```
In [81]: cnn_df
```

Out[81]:

	source	url	content	tokens	token_length	tokens_str
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	Chairman of the Joint Chiefs of Staff Gen. CQ ...	[chairman, joint, chiefs, staff, gen, cq, brow...	469	chairman joint chiefs staff gen cq brown monda...
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump has endorsed North Carolina Republican P...	[trump, endorsed, north, carolina, republican,...	104	trump endorsed north carolina republican chair...
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...	[senate, inching, closer, final, passage, 953,...	399	senate inching closer final passage 953 billio...
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Biden and King Abdullah II of Jordan met Monda...	[biden, king, abdullah, ii, jordan, met, monda...	642	biden king abdullah ii jordan met monday aimin...
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump on Monday asked the SupremeCourt to step...	[trump, monday, asked, supremecourt, step, cha...	567	trump monday asked supremecourt step charged d...
...	...	...	...	...	...	...
220	cnn	https://www.cnn.com/2024/02/15/politics/navy-f...	Members of Congress pressed the CEO of the nat...	[members, congress, pressed, ceo, nation's, la...	512	members congress pressed ceo nation's largest ...
221	cnn	https://www.cnn.com/2024/02/16/politics/kamala...	US Vice President Kamala Harris on Friday call...	[us, vice, president, kamala, harris, friday, ...	565	us vice president kamala harris friday called ...
222	cnn	https://www.cnn.com/2024/01/30/politics/trump-...	NewYork state Judge Arthur Engoron has the fut...	[newyork, state, judge, arthur, engoron, futur...	583	newyork state judge arthur engoron future trum...
223	cnn	https://www.cnn.com/2024/02/15/politics/border...	The acting deputy chief of the US Border Patro...	[acting, deputy, chief, us, border, patrol, jo...	101	acting deputy chief us border patrol joel mart...
224	cnn	https://www.cnn.com/2024/02/15/politics/former...	Special counsel David Weiss charged a former F...	[special, counsel, david, weiss, charged, fbi,...	456	special counsel david weiss charged fbi inform...

143 rows × 6 columns

In [82]:

```
fox_df
```

Out[82]:

	source	url	content	tokens	token_length	tokens_str
47	foxnews	https://www.foxnews.com/politics/biden-takes-j...	Biden takes jab at special counsel report with...	[biden, takes, jab, special, counsel, report, ...	473	biden takes jab special counsel report joke me...
48	foxnews	https://www.foxnews.com/politics/rfk-jr-apolog...	RFK Jr. drops surprise campaign ad during Supe...	[rfk, jr, drops, surprise, campaign, ad, super...	323	rfk jr drops surprise campaign ad super bowl a...
49	foxnews	https://www.foxnews.com/politics/bidens-upcomi...	Biden won't take cognitive test in physical ex...	[biden, wont, take, cognitive, test, physical,...	228	biden wont take cognitive test physical exam w...
50	foxnews	https://www.foxnews.com/politics/kamala-harris...	Marc Thiessen questions whether Biden is capab...	[marc, thiessen, questions, whether, biden, ca...	274	marc thiessen questions whether biden capable ...
51	foxnews	https://www.foxnews.com/politics/climate-activ...	Biden export suspension on liquefied natural g...	[biden, export, suspension, liquefied, natural...	451	biden export suspension liquefied natural gas ...
...	...	...	...	...	...	...
235	foxnews	https://www.foxnews.com/politics/house-republi...	Rep. Ronny Jackson demands Biden take cognitiv...	[rep, ronny, jackson, demands, biden, take, co...	464	rep ronny jackson demands biden take cognitive...
236	foxnews	https://www.foxnews.com/politics/gop-senators-...	Biden and the Democrat just do not care: Sen. ...	[biden, democrat, care, sen, ted, cruz, sen, t...	325	biden democrat care sen ted cruz sen ted cruz ...
237	foxnews	https://www.foxnews.com/politics/doj-defends-s...	Former US attorney discusses Special Counsel H...	[us, attorney, discusses, special, counsel, hu...	382	us attorney discusses special counsel hurs app...
238	foxnews	https://www.foxnews.com/politics/fox-news-poli...	Welcome to Fox News' Politics newsletter with ...	[welcome, news', politics, newsletter, latest,...	139	welcome news' politics newsletter latest polit...
239	foxnews	https://www.foxnews.com/politics/democrats-win...	Dems flipping NY House seat threatens GOP majo...	[dems, flipping, ny, house, seat, threatens, g...	545	dems flipping ny house seat threatens gop majo...

97 rows × 6 columns

In [83]:

```
# create our count text vectorizers
cnn_count_text_vectorizer = CountVectorizer(stop_words=list(sw), min_df=3, max_df=0.7)
cnn_count_text_vectors = cnn_count_text_vectorizer.fit_transform(cnn_df["tokens_str"])
print(cnn_count_text_vectors.shape)

fox_count_text_vectorizer = CountVectorizer(stop_words=list(sw), min_df=3, max_df=0.7)
fox_count_text_vectors = fox_count_text_vectorizer.fit_transform(fox_df["tokens_str"])
print(fox_count_text_vectors.shape)
```

```
(143, 3695)
(97, 2363)
```

In [84]:

```
# create our tf-idf text vectorizers
cnn_tfidf_text_vectorizer = TfidfVectorizer(stop_words=list(sw), min_df=3, max_df=0.7)
cnn_tfidf_text_vectors = cnn_tfidf_text_vectorizer.fit_transform(cnn_df['tokens_str'])
print(cnn_tfidf_text_vectors.shape)

fox_tfidf_text_vectorizer = TfidfVectorizer(stop_words=list(sw), min_df=3, max_df=0.7)
```

```
fox_tfidf_text_vectors = fox_tfidf_text_vectorizer.fit_transform(fox_df['tokens_str'])
print(fox_tfidf_text_vectors.shape)
```

```
(143, 3695)
(97, 2363)
```

Fitting a Non-Negative Matrix Factorization Model

5 Topics

In [85]:

```
# fit our NMF models
cnn_nmf_model = NMF(n_components=5, random_state=314)
cnn_W_nmf_matrix = cnn_nmf_model.fit_transform(cnn_tfidf_text_vectors)
cnn_H_nmf_matrix = cnn_nmf_model.components_

fox_nmf_model = NMF(n_components=5, random_state=315)
fox_W_nmf_matrix = fox_nmf_model.fit_transform(fox_tfidf_text_vectors)
fox_H_nmf_matrix = fox_nmf_model.components_
```

In [86]:

```
# assertion statements to ensure the document-topic and topic-feature matrices have the intended shapes
assert cnn_W_nmf_matrix.shape == (143, 5), f"Expected shape (143, 5), but got {cnn_W_nmf_matrix.shape}"
assert cnn_H_nmf_matrix.shape == (5, 3695), f"Expected shape (5, 3695), but got {cnn_H_nmf_matrix.shape}"
assert fox_W_nmf_matrix.shape == (97, 5), f"Expected shape (97, 5), but got {fox_W_nmf_matrix.shape}"
assert fox_H_nmf_matrix.shape == (5, 2363), f"Expected shape (5, 2363), but got {fox_H_nmf_matrix.shape}"
```

In [87]:

```
display_topics(cnn_nmf_model, cnn_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  trump (2.40)
  case (1.01)
  trial (0.86)
  willis (0.86)
  court (0.79)

Topic 01
  bill (1.53)
  aid (1.52)
  ukraine (1.51)
  senate (1.44)
  border (1.40)

Topic 02
  biden (1.82)
  hur (1.69)
  report (1.22)
  classified (1.11)
  documents (0.94)

Topic 03
  suozzi (1.99)
  democrat (1.29)
  pilip (0.99)
  santos (0.96)
  republican (0.84)

Topic 04
  nato (2.03)
  trump (1.57)
  us (0.88)
  russia (0.74)
  biden (0.63)
```

In [88]:

```
display_topics(fox_nmf_model, fox_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  aid (1.19)
  senate (1.03)
  border (1.03)
  bill (0.97)
  package (0.92)

Topic 01
  bobulinski (3.24)
  hunterbiden (2.35)
  2017 (1.10)
  business (1.07)
  hunter (1.05)

Topic 02
  hur (1.31)
  report (1.19)
  special (1.17)
  counsel (1.09)
  classified (1.03)

Topic 03
  trump (2.50)
  election (0.88)
  haley (0.70)
  republican (0.69)
  suozzi (0.67)

Topic 04
  manchin (5.36)
  romney (3.47)
  sen (2.41)
  mitt (2.36)
  running (2.18)
```

In [89]:
```
cnn_W_nmf_matrix.sum(axis=0)/cnn_W_nmf_matrix.sum()*100.0
```

Out[89]:
```
array([18.83568536, 21.55767503, 19.84341513, 17.61373797, 22.14948651])
```

In [90]:
```
fox_W_nmf_matrix.sum(axis=0)/fox_W_nmf_matrix.sum()*100.0
```

Out[90]:
```
array([20.55896756, 17.3733471 , 25.82227589, 26.73961564,  9.50579381])
```

4 Topics

In [91]:
```
# fit our NMF models 4
cnn_nmf_model4 = NMF(n_components=4, random_state=314)
cnn_W_nmf_matrix4 = cnn_nmf_model4.fit_transform(cnn_tfidf_text_vectors)
cnn_H_nmf_matrix4 = cnn_nmf_model4.components_

fox_nmf_model4 = NMF(n_components=4, random_state=315)
fox_W_nmf_matrix4 = fox_nmf_model4.fit_transform(fox_tfidf_text_vectors)
fox_H_nmf_matrix4 = fox_nmf_model4.components_
```

In [92]:
```
display_topics(cnn_nmf_model4, cnn_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  trump (2.49)
  case (0.86)
  trial (0.73)
  willis (0.70)
  election (0.69)

Topic 01
  ukraine (1.39)
  aid (1.24)
  bill (1.23)
  senate (1.14)
  border (1.10)

Topic 02
  biden (1.72)
  hur (1.38)
  report (1.02)
  classified (0.91)
  documents (0.76)

Topic 03
  suozzi (1.99)
  democrat (1.28)
  pilip (0.99)
  santos (0.96)
  republican (0.86)
```

In [93]:

```
display_topics(fox_nmf_model4, fox_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  aid (1.18)
  senate (1.03)
  border (1.02)
  bill (0.96)
  package (0.91)

Topic 01
  bobulinski (3.23)
  hunterbiden (2.35)
  2017 (1.10)
  business (1.07)
  hunter (1.05)

Topic 02
  hur (1.31)
  report (1.19)
  special (1.17)
  counsel (1.09)
  classified (1.04)

Topic 03
  trump (2.34)
  election (0.82)
  republican (0.68)
  haley (0.66)
  suozzi (0.61)
```

In [94]:

```
cnn_W_nmf_matrix4.sum(axis=0)/cnn_W_nmf_matrix4.sum()*100.0
```

Out[94]:

```
array([25.64844998, 28.67559657, 25.32015688, 20.35579657])
```

In [95]:

```
fox_W_nmf_matrix4.sum(axis=0)/fox_W_nmf_matrix4.sum()*100.0
```

Out[95]:

```
array([22.58918529, 18.70581594, 27.65188173, 31.05311704])
```

Fitting an LSA Model

5 Topics

In [96]:

```
# fit our LSA models
cnn_svd_model = TruncatedSVD(n_components=5, random_state=320)
cnn_W_svd_matrix = cnn_svd_model.fit_transform(cnn_tfidf_text_vectors)
```

```
cnn_H_svd_matrix = cnn_svd_model.components_

fox_svd_model = TruncatedSVD(n_components=5, random_state=321)
fox_W_svd_matrix = fox_svd_model.fit_transform(fox_tfidf_text_vectors)
fox_H_svd_matrix = fox_svd_model.components_
```

In [97]:

```
# assertion statements to ensure the document-topic and topic-feature matrices have the intended shapes
assert cnn_W_svd_matrix.shape == (143, 5), f"Expected shape (143, 5), but got {cnn_W_svd_matrix.shape}"
assert cnn_H_svd_matrix.shape == (5, 3695), f"Expected shape (5, 3695), but got {cnn_H_svd_matrix.shape}"
assert fox_W_svd_matrix.shape == (97, 5), f"Expected shape (97, 5), but got {fox_W_svd_matrix.shape}"
assert fox_H_svd_matrix.shape == (5, 2363), f"Expected shape (5, 2363), but got {fox_H_svd_matrix.shape}"
```

In [98]:

```
display_topics(cnn_svd_model, cnn_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  trump (1.31)
  biden (0.58)
  republican (0.45)
  election (0.40)
  case (0.34)

Topic 01
  ukraine (6.30)
  aid (5.76)
  bill (5.68)
  senate (5.50)
  border (5.25)

Topic 02
  biden (5.55)
  hur (5.28)
  report (3.76)
  classified (3.44)
  documents (2.90)

Topic 03
  suozzi (16.62)
  democrat (9.45)
  pilip (8.20)
  santos (7.99)
  newyork (6.37)

Topic 04
  nato (24.78)
  haley (11.60)
  trump (10.63)
  suozzi (9.25)
  biden (8.63)
```

In [99]:

```
display_topics(fox_svd_model, fox_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  trump (0.78)
  republican (0.46)
  special (0.39)
  senate (0.39)
  border (0.37)

Topic 01
  bobulinski (19.09)
  hunterbiden (14.55)
  business (6.85)
  2017 (6.79)
  hunter (6.55)

Topic 02
  hur (16.67)
  special (15.21)
  report (14.67)
  counsel (14.02)
  classified (13.41)

Topic 03
  trump (46.93)
  election (16.40)
  haley (16.33)
  willis (13.47)
  newyork (13.04)

Topic 04
  manchin (14.81)
  romney (9.61)
  sen (6.66)
  mitt (6.53)
  running (5.92)
```

In [100…

```
cnn_W_svd_matrix.sum(axis=0)/cnn_W_svd_matrix.sum()*100.0
```

Out[100…

```
array([86.27215675,  6.2283855 ,  6.9115943 , -0.11539489,  0.70325833])
```

In [101…

```
fox_W_svd_matrix.sum(axis=0)/fox_W_svd_matrix.sum()*100.0
```

Out[101…

```
array([90.21771346,  3.75501037, -0.76965611,  0.90394909,  5.8929832 ])
```

4 Topics

In [102…

```
# fit our LSA models 4
cnn_svd_model4 = TruncatedSVD(n_components=4, random_state=320)
cnn_W_svd_matrix4 = cnn_svd_model4.fit_transform(cnn_tfidf_text_vectors)
cnn_H_svd_matrix4 = cnn_svd_model4.components_

fox_svd_model4 = TruncatedSVD(n_components=4, random_state=321)
fox_W_svd_matrix4 = fox_svd_model4.fit_transform(fox_tfidf_text_vectors)
fox_H_svd_matrix4 = fox_svd_model4.components_
```

In [103…

```
display_topics(cnn_svd_model4, cnn_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  trump (1.31)
  biden (0.58)
  republican (0.45)
  election (0.40)
  case (0.34)

Topic 01
  ukraine (6.31)
  aid (5.77)
  bill (5.69)
  senate (5.51)
  border (5.25)

Topic 02
  biden (5.55)
  hur (5.29)
  report (3.76)
  classified (3.45)
  documents (2.90)

Topic 03
  suozzi (16.11)
  democrat (9.15)
  pilip (7.95)
  santos (7.74)
  newyork (6.22)
```

In [104…

```
display_topics(fox_svd_model4, fox_tfidf_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  trump (0.78)
  republican (0.46)
  special (0.39)
  senate (0.39)
  border (0.37)

Topic 01
  bobulinski (19.36)
  hunterbiden (14.74)
  business (6.94)
  2017 (6.88)
  hunter (6.63)

Topic 02
  hur (16.63)
  special (15.14)
  report (14.61)
  counsel (13.99)
  classified (13.41)

Topic 03
  trump (48.32)
  election (16.77)
  haley (16.74)
  willis (14.04)
  newyork (13.50)
```

In [105…

```
cnn_W_svd_matrix4.sum(axis=0)/cnn_W_svd_matrix4.sum()*100.0
```

Out[105…

```
array([ 8.68549898e+01,  6.26796756e+00,  6.96257431e+00, -8.55316871e-02])
```

In [106…

```
fox_W_svd_matrix4.sum(axis=0)/fox_W_svd_matrix4.sum()*100.0
```

Out[106…

```
array([95.84248683,  3.98084627, -0.83215167,  1.00881857])
```

Fitting an LDA Model

5 Topics

In [107…

```
# fit our LDA models
cnn_lda_model = LatentDirichletAllocation(n_components=5, random_state=40)
cnn_W_lda_matrix = cnn_lda_model.fit_transform(cnn_count_text_vectors)
```

```
cnn_H_lda_matrix = cnn_lda_model.components_

fox_lda_model = LatentDirichletAllocation(n_components=5, random_state=41)
fox_W_lda_matrix = fox_lda_model.fit_transform(fox_count_text_vectors)
fox_H_lda_matrix = fox_lda_model.components_
```

In [108…

```
# assertion statements to ensure the document-topic and topic-feature matrices have the intended shapes
assert cnn_W_lda_matrix.shape == (143, 5), f"Expected shape (143, 5), but got {cnn_W_lda_matrix.shape}"
assert cnn_H_lda_matrix.shape == (5, 3695), f"Expected shape (5, 3695), but got {cnn_H_lda_matrix.shape}"
assert fox_W_lda_matrix.shape == (97, 5), f"Expected shape (97, 5), but got {fox_W_lda_matrix.shape}"
assert fox_H_lda_matrix.shape == (5, 2363), f"Expected shape (5, 2363), but got {fox_H_lda_matrix.shape}"
```

In [109…

```
display_topics(cnn_lda_model, cnn_count_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  democrat (2.27)
  republican (2.26)
  suozzi (1.43)
  vote (1.19)
  gop (1.14)

Topic 01
  biden (3.48)
  report (0.94)
  hur (0.89)
  republican (0.68)
  trump (0.65)

Topic 02
  willis (2.39)
  wade (1.49)
  trump (1.14)
  senate (1.10)
  democrat (1.03)

Topic 03
  trump (5.73)
  election (1.28)
  case (1.14)
  court (0.83)
  trial (0.73)

Topic 04
  us (1.57)
  ukraine (1.14)
  republican (1.08)
  aid (0.82)
  nato (0.82)
```

In [110…

```
display_topics(fox_lda_model, fox_count_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  report (1.63)
  special (1.61)
  hur (1.53)
  counsel (1.50)
  memory (1.30)

Topic 01
  security (1.03)
  trump (0.97)
  ukraine (0.81)
  national (0.80)
  willis (0.79)

Topic 02
  border (1.51)
  republican (1.36)
  senate (1.30)
  vote (0.95)
  mayorkas (0.83)

Topic 03
  bobulinski (1.62)
  hunterbiden (1.47)
  business (1.11)
  chinese (0.71)
  2017 (0.66)

Topic 04
  trump (2.90)
  republican (1.49)
  democrat (1.24)
  election (0.99)
  campaign (0.70)
```

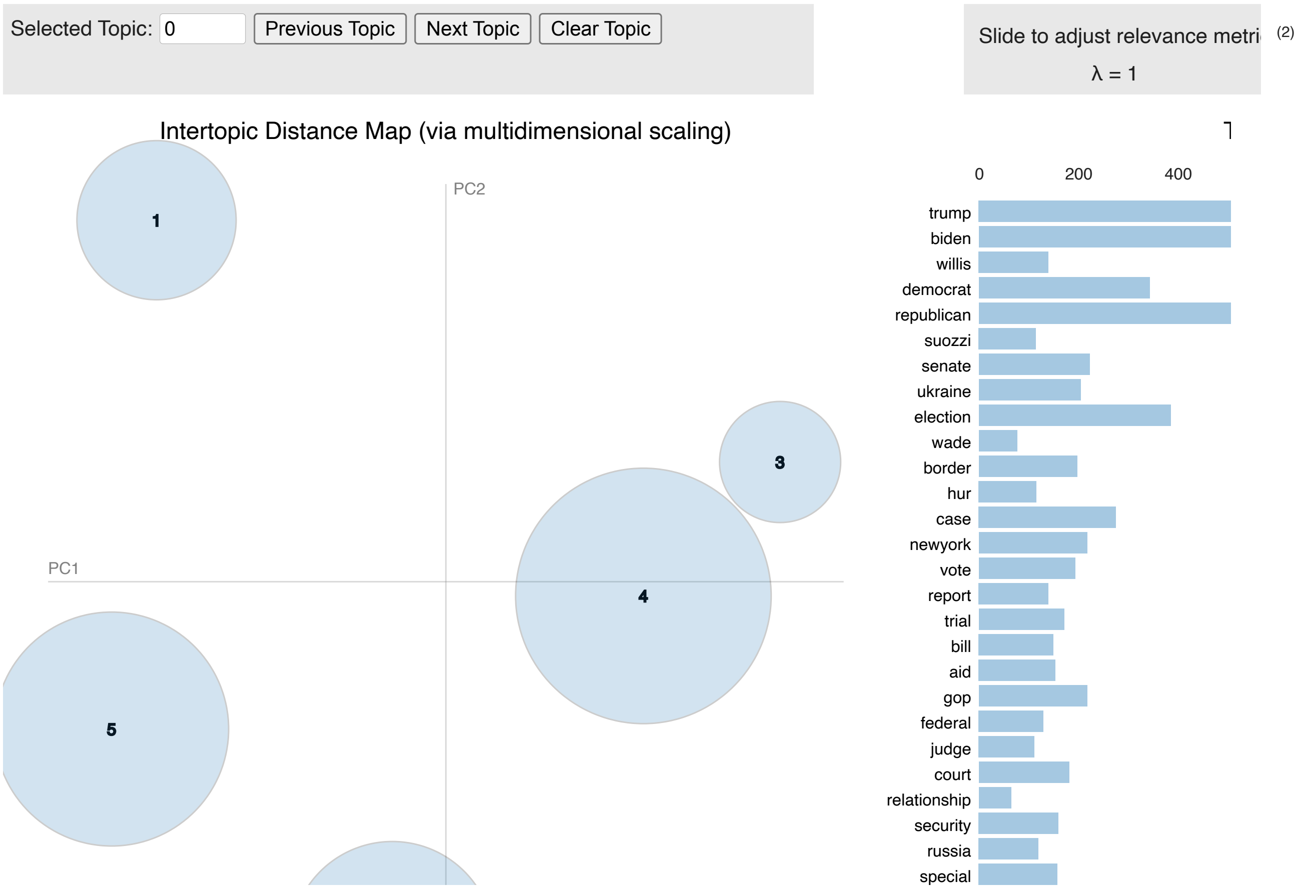
In [111…

```
# prepare our models for display
cnn_lda_display = pyLDAvis.lda_model.prepare(cnn_lda_model, cnn_count_text_vectors, cnn_count_text_vectori
fox_lda_display = pyLDAvis.lda_model.prepare(fox_lda_model, fox_count_text_vectors, fox_count_text_vectori
```

In [112…

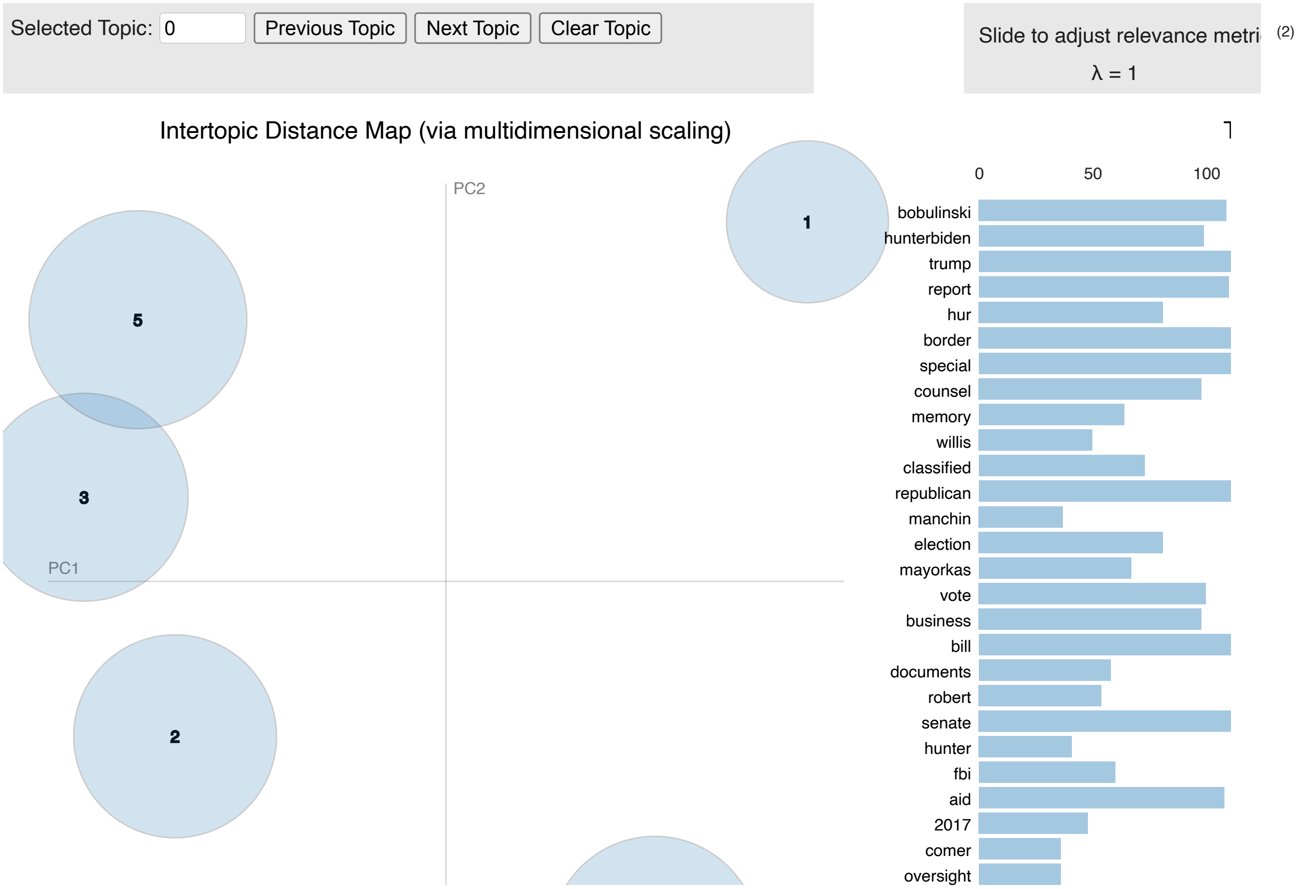
```
pyLDAvis.display(cnn_lda_display)
```

```
Out[112…
```



```
In [113…  pyLDAvis.display(fox_lda_display)
```

```
Out[113…
```



4 Topics

In [114…

```
# fit our LDA models 4
cnn_lda_model4 = LatentDirichletAllocation(n_components=4, random_state=40)
cnn_W_lda_matrix4 = cnn_lda_model4.fit_transform(cnn_count_text_vectors)
cnn_H_lda_matrix4 = cnn_lda_model4.components_

fox_lda_model4 = LatentDirichletAllocation(n_components=4, random_state=41)
fox_W_lda_matrix4 = fox_lda_model4.fit_transform(fox_count_text_vectors)
fox_H_lda_matrix4 = fox_lda_model4.components_
```

In [115…

```
display_topics(cnn_lda_model4, cnn_count_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  republican (2.20)
  democrat (1.27)
  border (1.16)
  senate (1.05)
  gop (0.93)

Topic 01
  biden (2.65)
  us (1.09)
  report (0.69)
  hur (0.60)
  white (0.54)

Topic 02
  willis (2.27)
  wade (1.31)
  trump (1.18)
  case (1.03)
  relationship (0.85)

Topic 03
  trump (5.56)
  election (1.21)
  case (0.89)
  court (0.76)
  trial (0.63)
```

In [116…

```
display_topics(fox_lda_model4, fox_count_text_vectorizer.get_feature_names_out())
```

```
Topic 00
  report (1.40)
  special (1.33)
  hur (1.18)
  counsel (1.16)
  memory (1.11)

Topic 01
  trump (2.42)
  republican (0.97)
  democrat (0.74)
  campaign (0.67)
  security (0.66)

Topic 02
  border (1.30)
  senate (1.26)
  republican (1.25)
  bill (1.01)
  aid (0.87)

Topic 03
  bobulinski (1.47)
  hunterbiden (1.30)
  business (1.00)
  chinese (0.70)
  fbi (0.62)
```

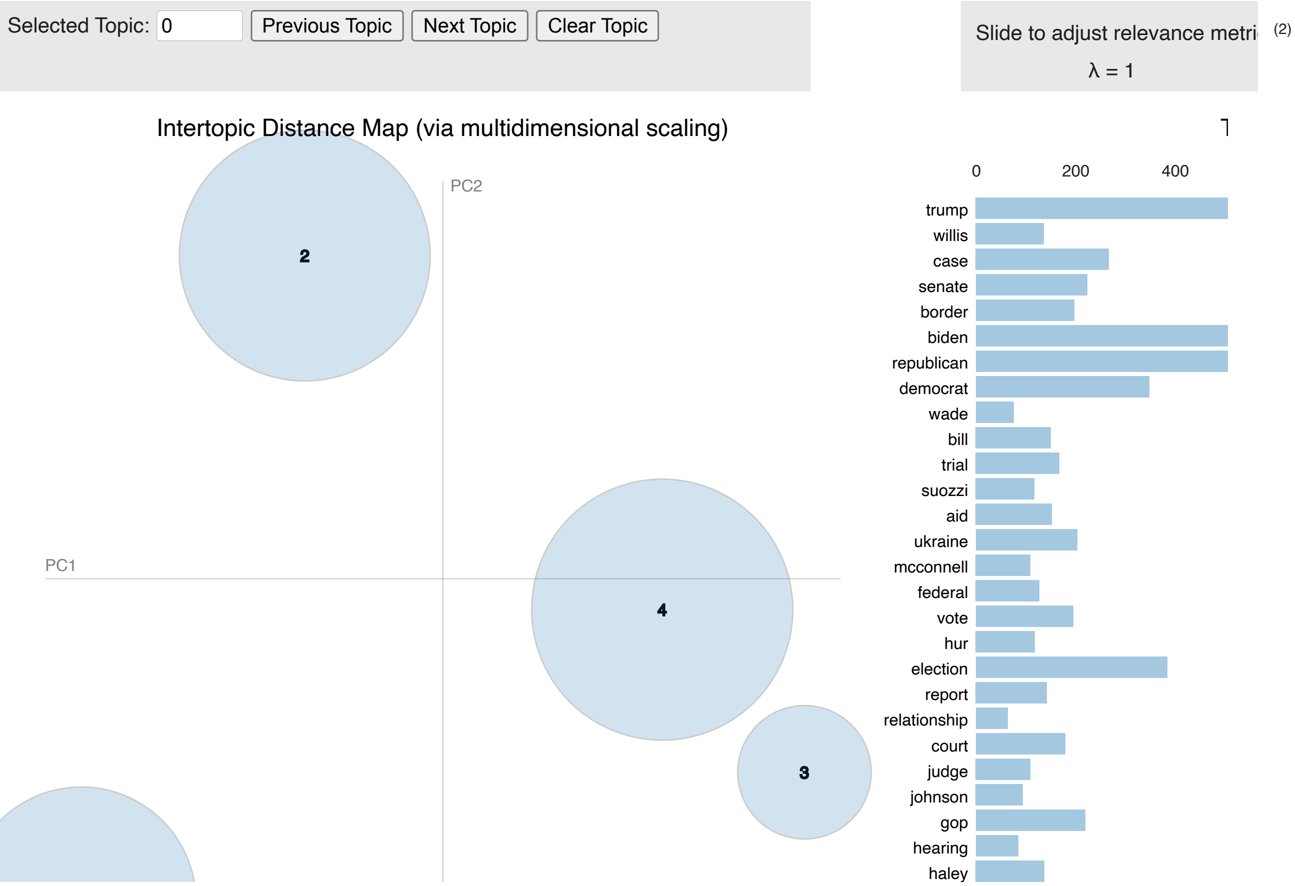
In [117…

```
cnn_lda_display4 = pyLDAvis.lda_model.prepare(cnn_lda_model4, cnn_count_text_vectors, cnn_count_text_vecto
fox_lda_display4 = pyLDAvis.lda_model.prepare(fox_lda_model4, fox_count_text_vectors, fox_count_text_vecto
```

In [118…

```
pyLDAvis.display(cnn_lda_display4)
```

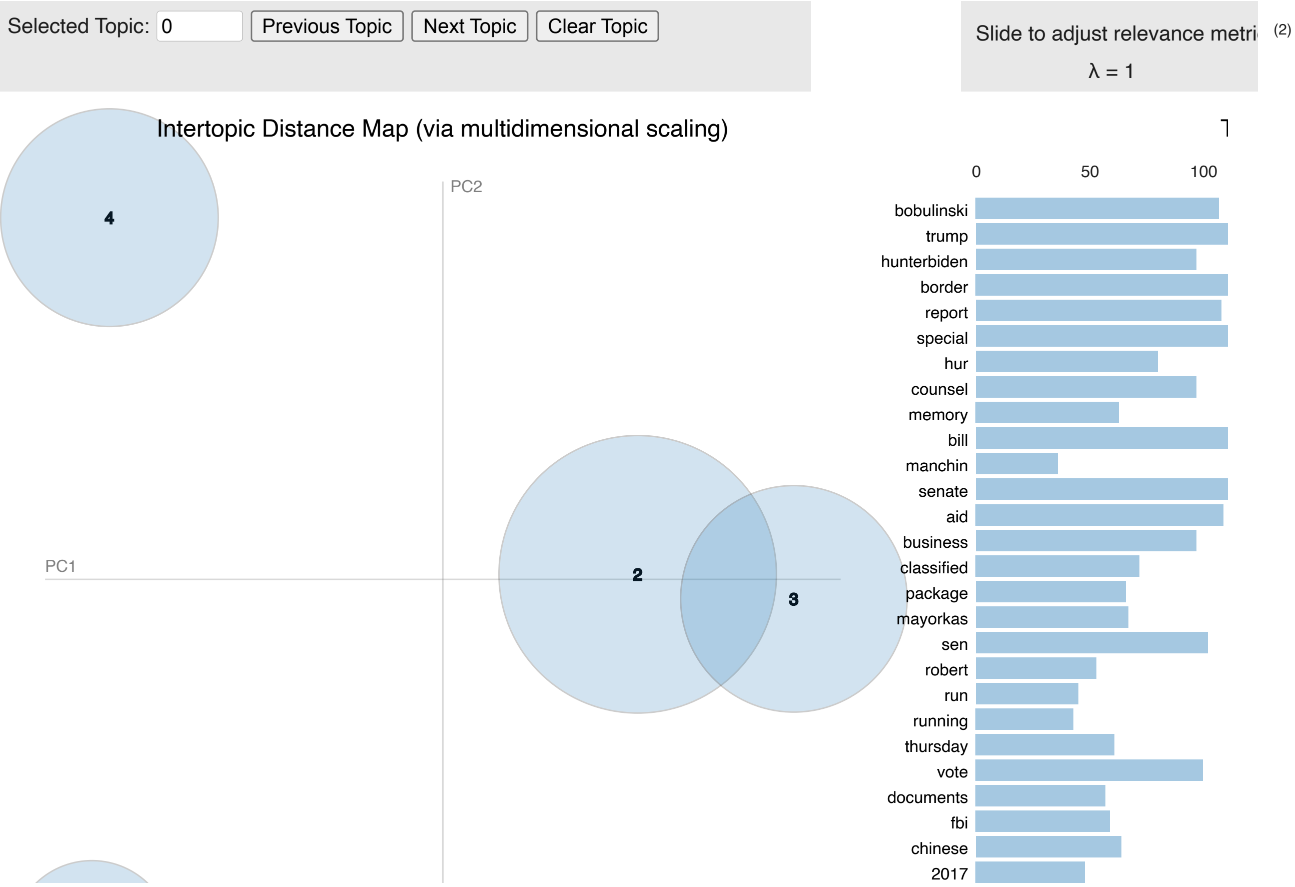
Out[118…



In [119…

```
pyLDAvis.display(fox_lda_display4)
```

Out[119…



Sentiment Analysis

In [120…

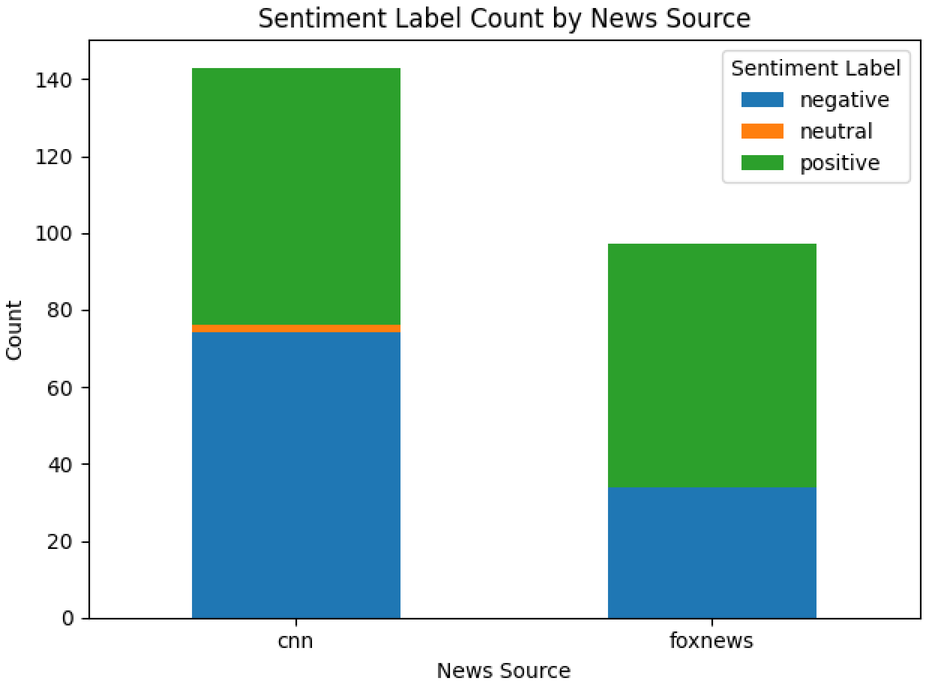
```
sid = SentimentIntensityAnalyzer()

def get_sentiment_scores(text):

    text_str = ' '.join(text)
    return sid.polarity_scores(text_str)

# get sentiment scores for each news article
df['sentiment_scores'] = df['tokens'].apply(get_sentiment_scores)

# Extract compound sentiment scores (normalized score between -1 (most negative) and +1 (most positive))
df['compound_sentiment'] = df['sentiment_scores'].apply(lambda x: x['compound'])

threshold = 0.05

df['sentiment_label'] = df['compound_sentiment'].apply(lambda score: 'positive' if score > threshold else

df.head()
```

Out[120…

	source	url	content	tokens	token_length	sentiment_scores	c
0	cnn	https://www.cnn.com/2024/02/12/politics/cq-bro...	Chairman of the Joint Chiefs of Staff Gen. CQ ...	[chairman, joint, chiefs, staff, gen, cq, brow...	469	{'neg': 0.127, 'neu': 0.775, 'pos': 0.098, 'co...	
1	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump has endorsed North Carolina Republican P...	[trump, endorsed, north, carolina, republican,...	104	{'neg': 0.029, 'neu': 0.68, 'pos': 0.291, 'com...	
2	cnn	https://www.cnn.com/2024/02/12/politics/senate...	The Senate is inching closer to final passage ...	[senate, inching, closer, final, passage, 953,...	399	{'neg': 0.105, 'neu': 0.721, 'pos': 0.174, 'co...	
3	cnn	https://www.cnn.com/2024/02/12/politics/bidens...	Biden and King Abdullah II of Jordan met Monda...	[biden, king, abdullah, ii, jordan, met, monda...	642	{'neg': 0.148, 'neu': 0.766, 'pos': 0.086, 'co...	
4	cnn	https://www.cnn.com/2024/02/12/politics/trump-...	Trump on Monday asked the SupremeCourt to step...	[trump, monday, asked, supremecourt, step, cha...	567	{'neg': 0.168, 'neu': 0.761, 'pos': 0.071, 'co...	

In [121…

```
# Group by source and sentiment label and count occurrences
sentiment_counts = df.groupby(['source', 'sentiment_label']).size().unstack(fill_value=0)

sentiment_counts.plot(kind='bar', stacked=True)
plt.title('Sentiment Label Count by News Source')
plt.xlabel('News Source')
plt.ylabel('Count')
plt.xticks(rotation=0)
plt.legend(title='Sentiment Label')
plt.tight_layout()
plt.show()
```



Binary Classification - Source Prediction

Predicting News Sources with Random Forest Classifier

In [122…

```
# Define X and y
X = df['tokens']
y = df['source']

# Convert list of tokens into strings and remove stop words
X_str = X.apply(lambda x: ' '.join([token for token in x if token.lower() not in sw]))

tfidf_vectorizer = TfidfVectorizer()
X_tfidf = tfidf_vectorizer.fit_transform(X_str)
X_train, X_test, y_train, y_test = train_test_split(X_tfidf, y, test_size=0.2, random_state=42)
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

# Model Evaluation
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.8333333333333334
Classification Report:
              precision    recall  f1-score   support

         cnn       0.76      1.00      0.87        26
     foxnews       1.00      0.64      0.78        22

    accuracy                           0.83        48
   macro avg       0.88      0.82      0.82        48
weighted avg       0.87      0.83      0.83        48
```

In [123…

```
feature_importances = model.feature_importances_
feature_names = tfidf_vectorizer.get_feature_names_out()
feature_importance_dict = dict(zip(feature_names, feature_importances))
sorted_feature_importances = sorted(feature_importance_dict.items(), key=lambda x: x[1], reverse=True)

top_n = 20
print(f"Top {top_n} features and their importances:")
```

```
for feature, importance in sorted_feature_importances[:top_n]:
    print(f"Feature: {feature}, Importance: {importance}")
```

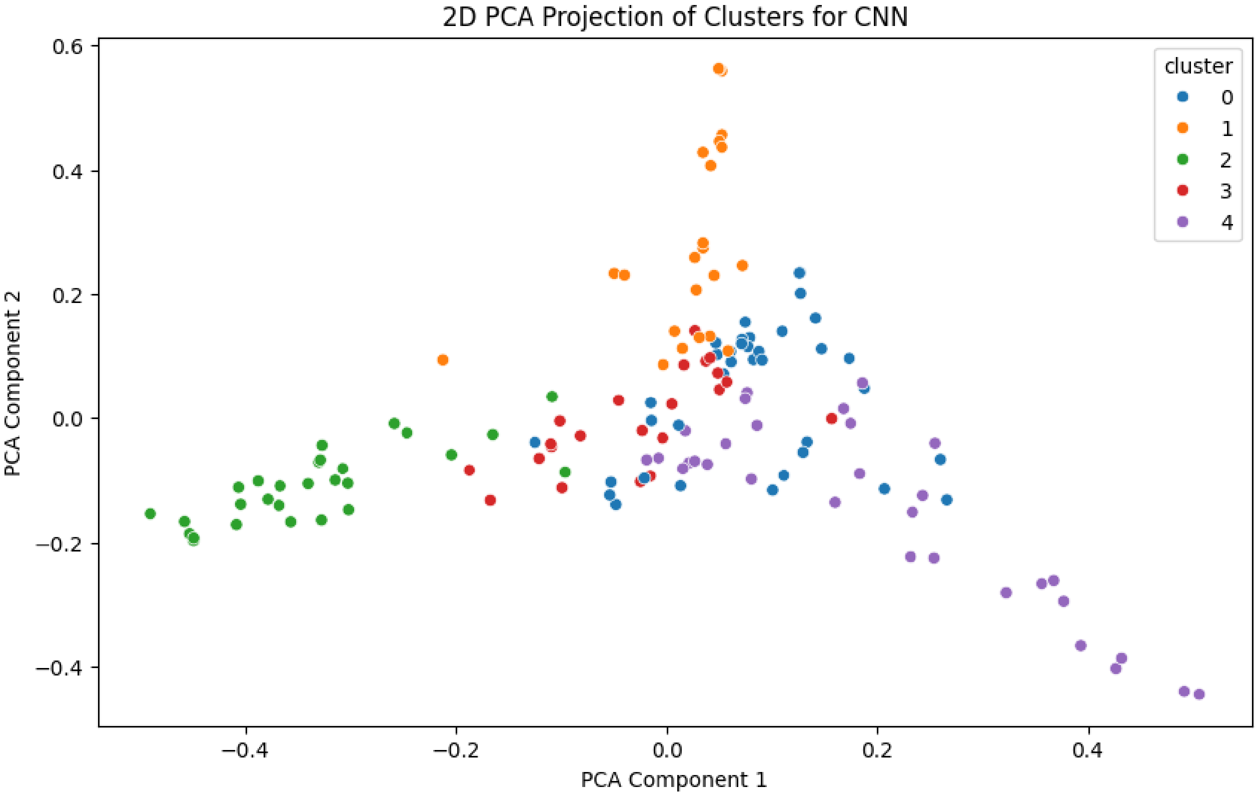
```
Top 20 features and their importances:
Feature: the, Importance: 0.019457762723102442
Feature: we, Importance: 0.010278404387749384
Feature: please, Importance: 0.01020224845438076
Feature: biden, Importance: 0.0094910269548895
Feature: feb, Importance: 0.008807746745386077
Feature: dont, Importance: 0.008789921490812182
Feature: hill, Importance: 0.007784379245635765
Feature: content, Importance: 0.006711793845796507
Feature: dc, Importance: 0.0063161376667195445
Feature: 2024, Importance: 0.006175476550169621
Feature: latest, Importance: 0.005844767115020217
Feature: it, Importance: 0.0055366298045883555
Feature: valid, Importance: 0.005125129609628121
Feature: bidens, Importance: 0.004862530949031991
Feature: related, Importance: 0.004729150903611669
Feature: one, Importance: 0.004721953446508506
Feature: including, Importance: 0.004450905004907464
Feature: reporter, Importance: 0.004431897923624123
Feature: access, Importance: 0.00435665666819693
Feature: plus, Importance: 0.004098933712122046
```

Clustering

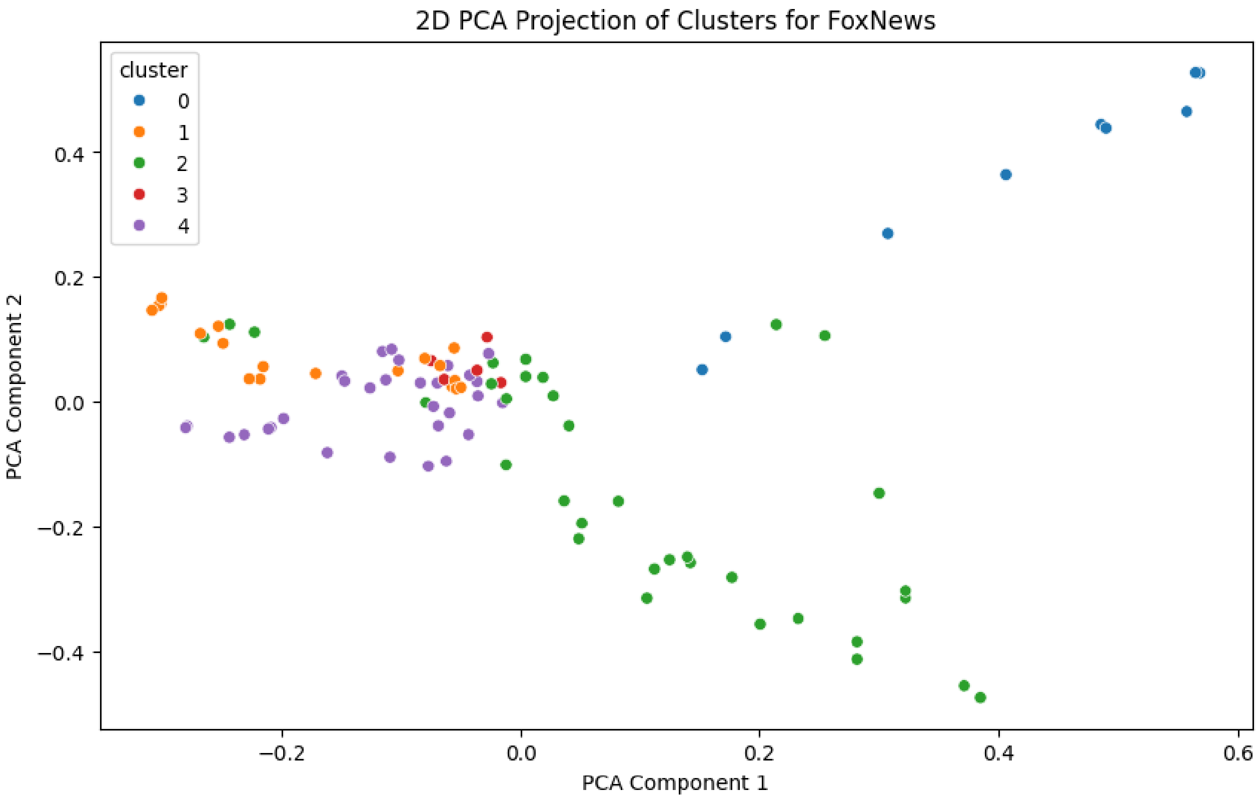
In [124…

```
def cluster_and_plot(df, source_name):

    df['text'] = df['tokens'].apply(lambda x: ' '.join(x))
    tfidf_vectorizer = TfidfVectorizer()
    X_tfidf = tfidf_vectorizer.fit_transform(df['text'])

    # K-means clustering
    k = 5
    kmeans = KMeans(n_clusters=k, random_state=42)
    clusters = kmeans.fit_predict(X_tfidf)
    df['cluster'] = clusters

    # Print the top words per cluster
    print(f"Top words per cluster for {source_name}:")
    order_centroids = kmeans.cluster_centers_.argsort()[:, ::-1]
    terms = tfidf_vectorizer.get_feature_names_out()
    for i in range(k):
        print(f"Cluster {i}: ", end='')
        for ind in order_centroids[i, :10]:
            print(f'{terms[ind]}', end=', ')
        print()

    # Reduce dimensions to 2D using PCA
    pca = PCA(n_components=2)
    X_pca = pca.fit_transform(X_tfidf.toarray())

    # Add PCA components to DataFrame
    df['pca1'] = X_pca[:, 0]
    df['pca2'] = X_pca[:, 1]

    plt.figure(figsize=(10, 6))
    sns.scatterplot(data=df, x='pca1', y='pca2', hue='cluster', palette='tab10', legend='full')
    plt.title(f'2D PCA Projection of Clusters for {source_name}')
    plt.xlabel('PCA Component 1')
    plt.ylabel('PCA Component 2')
    plt.show()

# CNN articles
df_cnn = df[df['source'] == 'cnn']
cluster_and_plot(df_cnn, 'CNN')

# FoxNews articles
df_fox = df[df['source'] == 'foxnews']
cluster_and_plot(df_fox, 'FoxNews')
```

```
Top words per cluster for CNN:
Cluster 0: nato, trump, biden, us, russia, ukraine, austin, intelligence, netanyahu, defense,
Cluster 1: biden, hur, report, classified, fbi, documents, counsel, president, special, bobulinski,
Cluster 2: trump, case, willis, trial, court, supremecourt, election, newyork, judge, wade,
Cluster 3: trump, haley, kennedy, rnc, whatley, southcarolina, republican, biden, border, election,
Cluster 4: senate, republican, border, aid, house, suozzi, bill, democrat, mcconnell, ukraine,
```



```
Top words per cluster for FoxNews:
Cluster 0: bobulinski, hunterbiden, biden, hunter, business, 2017, cefc, drug, cocaine, energy,
Cluster 1: border, house, senate, aid, mayorkas, bill, johnson, package, security, republican,
Cluster 2: biden, hur, president, special, report, counsel, classified, memory, documents, house,
Cluster 3: hamas, israel, israeli, palestinian, oct, wray, sexual, hayes, twostate, solution,
Cluster 4: trump, republican, manchin, election, willis, haley, democrat, ramaswamy, suozzi, campaign,
```


2D PCA Projection of Clusters for FoxNews
cluster
● 0
● 1
● 2
● 3
● 4
0.4
0.2
0.0
−0.2
−0.4
PCA Component 2
−0.2
0.0
0.2
0.4
0.6
PCA Component 1