2/4/2024

M Junaid Farooq

20f-0427

BCS-8A

Web Scrapper

Web Scrapper for **AltNews:**

**Code:**

import requests  
from bs4 import BeautifulSoup  
import pandas as pd  
  
  
def get\_image\_source(img):  
 if img.get('src'):  
 return img['src']  
 elif img.get('data-src'):  
 return img['data-src']  
 return None  
  
  
def extract\_videos(soup):  
 videos = []  
 for video in soup.find\_all('video'):  
 if video.get('src'):  
 videos.append(video['src'])  
 for source in video.find\_all('source'):  
 if source.get('src'):  
 videos.append(source['src'])  
 for iframe in soup.find\_all('iframe'):  
 if iframe.get('src'):  
 videos.append(iframe['src'])  
 return videos  
  
  
url = 'https://www.altnews.in'  
  
response = requests.get(url)  
soup = BeautifulSoup(response.text, 'html.parser')  
  
# Images  
images = soup.find\_all('img')  
image\_data = [{'src': get\_image\_source(img), 'alt': img.get(  
 'alt', '')} for img in images if get\_image\_source(img)]  
  
# Text  
text\_tags = ['p', 'h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'span', 'div']  
text\_data = [tag.text for tag in soup.find\_all(text\_tags)]  
text\_data = list(set(text\_data))  
text\_data = [text for text in text\_data if text.strip()]  
  
# Videos  
video\_data = extract\_videos(soup)  
  
df\_images = pd.DataFrame(image\_data)  
df\_images.to\_csv('images\_data.csv', index=False)  
  
df\_text = pd.DataFrame(text\_data, columns=['text'])  
df\_text.to\_csv('text\_data.csv', index=False)  
  
df\_videos = pd.DataFrame(video\_data, columns=['video\_src'])  
df\_videos.to\_csv('videos\_data.csv', index=False)  
  
print("Completed data scraping and saving results in texts, videos, and images.")

Web Scrapper for **Mastodon :**

**Code:**

import requests  
from bs4 import BeautifulSoup  
import pandas as pd  
  
  
def get\_image\_source(img):  
 if img.get('src'):  
 return img['src']  
 elif img.get('data-src'):  
 return img['data-src']  
 return None  
  
  
def extract\_videos(soup):  
 videos = []  
 for video in soup.find\_all('video'):  
 if video.get('src'):  
 videos.append(video['src'])  
 for source in video.find\_all('source'):  
 if source.get('src'):  
 videos.append(source['src'])  
 for iframe in soup.find\_all('iframe'):  
 if iframe.get('src'):  
 videos.append(iframe['src'])  
 return videos  
  
  
def extract\_links(soup):  
 links = []  
 for link in soup.find\_all('a'):  
 url = link.get('href')  
 text = link.text.strip()  
 if url:   
 links.append({'url': url, 'text': text})  
 return links  
  
  
url = 'https://mastodon.social/explore'  
  
response = requests.get(url)  
soup = BeautifulSoup(response.text, 'html.parser')  
  
# Images  
image\_data = [{'src': get\_image\_source(img), 'alt': img.get(  
 'alt', '')} for img in soup.find\_all('img') if get\_image\_source(img)]  
  
# Text  
text\_data = [tag.text for tag in soup.find\_all(  
 ['p', 'h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'span', 'div']) if tag.text.strip()]  
  
# Videos  
video\_data = extract\_videos(soup)  
  
# Links  
link\_data = extract\_links(soup)  
  
# Saving data to CSV  
pd.DataFrame(image\_data).to\_csv('images\_data.csv', index=False)  
pd.DataFrame(text\_data, columns=['text']).to\_csv('text\_data.csv', index=False)  
pd.DataFrame(video\_data, columns=['video\_src']).to\_csv(  
 'videos\_data.csv', index=False)  
pd.DataFrame(link\_data).to\_csv('links\_data.csv', index=False)  
  
print("Completed data scraping and saving results in texts, videos, and images.")

Web Scrapper for **Politifact:**

**Code:**

import requests  
from bs4 import BeautifulSoup  
import pandas as pd  
  
  
def get\_image\_source(img):  
 if img.get('src'):  
 return img['src']  
 elif img.get('data-src'):  
 return img['data-src']  
 return None  
  
  
def extract\_videos(soup):  
 videos = []  
 for video in soup.find\_all('video'):  
 if video.get('src'):  
 videos.append(video['src'])  
 for source in video.find\_all('source'):  
 if source.get('src'):  
 videos.append(source['src'])  
 for iframe in soup.find\_all('iframe'):  
 if iframe.get('src'):  
 videos.append(iframe['src'])  
 return videos  
  
  
url = 'https://www.politifact.com'  
  
response = requests.get(url)  
soup = BeautifulSoup(response.text, 'html.parser')  
  
# Images  
images = soup.find\_all('img')  
image\_data = [{'src': get\_image\_source(img), 'alt': img.get(  
 'alt', '')} for img in images if get\_image\_source(img)]  
  
# Text  
text\_tags = ['p', 'h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'span', 'div']  
text\_data = [tag.text for tag in soup.find\_all(text\_tags)]  
text\_data = list(set(text\_data))  
text\_data = [text for text in text\_data if text.strip()]  
  
# Videos  
video\_data = extract\_videos(soup)  
  
df\_images = pd.DataFrame(image\_data)  
df\_images.to\_csv('images\_data.csv', index=False)  
  
df\_text = pd.DataFrame(text\_data, columns=['text'])  
df\_text.to\_csv('text\_data.csv', index=False)  
  
df\_videos = pd.DataFrame(video\_data, columns=['video\_src'])  
df\_videos.to\_csv('videos\_data.csv', index=False)  
  
print("Completed data scraping and saving results in texts, videos, and images.")

**Blog Link:**

**https://medium.com/@junaidfarooq0427/navigating-the-web-scraping-maze-a-tale-of-three-distinct-challenges-9262b8eac95d**