

ADS509_CarrA_Final_Project_03_Query_v1

June 18, 2023

1 509 Final Project

This notebook reads in a CSV file that contain an attribute with raw HTML, then parses the article text using BeautifulSoup. A copy of the full DF is written to a CSV file for sharing with all collaborators in GitHub.

1.1 Globally import libraries

```
[1]: import numpy as np
import pandas as pd
import pymysql as mysql
import matplotlib.pyplot as plt
import os
import shutil
import re
import logging
import time
import datetime as dt
import zipfile
import requests
from bs4 import BeautifulSoup
import datetime
import re
import regex as rex
from collections import defaultdict, Counter
import random
import json
#import mysql.connector

# Set pandas global options
pd.options.display.max_rows = 17
```

1.2 Upload data from CSV

```
[2]: '''Dir nav citation:
https://softhints.com/python-change-directory-parent/'''
curr_dir = os.path.abspath(os.curdir)
print(curr_dir)
```

```
os.chdir("..")
up1_dir = os.path.abspath(os.curdir)
print(up1_dir)
```

C:\Users\acarr\Documents\GitHub\ADS509_Final_project\deliverables
C:\Users\acarr\Documents\GitHub\ADS509_Final_project

```
[3]: # change `data_location` to the location of the folder on your machine.
data_r_location = 'data_restricted'
data_location = 'data'

file_in_name = 'data_raw_amc.csv'
file_out_name = 'data_parsed_amc.csv'

file_in_path01 = os.path.join(up1_dir, data_r_location, file_in_name)
file_out_path = os.path.join(up1_dir, data_location, file_out_name)

print(f'CSV file in path: {file_in_path01}')
print(f'CSV file out path: {file_out_path}')
```

CSV file in path: C:\Users\acarr\Documents\GitHub\ADS509_Final_project\data_restricted\data_raw_amc.csv

CSV file out path:

C:\Users\acarr\Documents\GitHub\ADS509_Final_project\data\data_parsed_amc.csv

```
[4]: slct_tbl_full_df01 = pd.read_csv(file_in_path01)
display(slct_tbl_full_df01.head())
```

	text_id	source_name	author	\
0	1	CNN	Clare Foran,Nicky Robertson	
1	2	Fox News	Paul Steinhauser	
2	3	Fox News	Greg Wehner	
3	4	Fox News	Michael Ruiz	
4	5	Fox News	Brooke Singman	

	title	\
0	Senate races to avert default but vote timing ...	
1	First on Fox: Pro-Tim Scott super PAC launches...	
2	Pennsylvania bus driver allegedly used duct ta...	
3	Bob Lee murder: Cash App founder seen with sus...	
4	Senate GOP demands answers on security clearan...	

	url	publish_date	\
0	https://www.cnn.com/2023/06/01/politics/senate...	2023-06-01T09:00:40Z	
1	https://www.foxnews.com/politics/pro-scott-sup...	2023-06-01T16:12:56Z	
2	https://www.foxnews.com/us/pennsylvania-bus-dr...	2023-05-31T00:21:20Z	
3	https://www.foxnews.com/us/bob-lee-murder-cash...	2023-05-31T20:30:12Z	
4	https://www.foxnews.com/politics/senate-gop-de...	2023-05-31T22:11:38Z	

	article_text	content
0	<!DOCTYPE html>\n<html data-layout-uri="cms.cn...	NaN
1	<!DOCTYPE html>\n<html data-n-head="%7B%22lang...	NaN
2	<!DOCTYPE html>\n<html data-n-head="%7B%22lang...	NaN
3	<!DOCTYPE html>\n<html data-n-head="%7B%22lang...	NaN
4	<!DOCTYPE html>\n<html data-n-head="%7B%22lang...	NaN

1.3 Extract article data

1.3.1 Check for missing articles

```
[5]: count_nan = slct_tbl_full_df01['article_text'].isnull().sum()

# printing the number of values present
# in the column
print('Number of NaN values present: ' + str(count_nan))
```

Number of NaN values present: 0

1.3.2 Parse the article text from the column with raw HTML

```
[6]: slct_tbl_full_df02 = slct_tbl_full_df01.copy()
#slct_tbl_full_df02 = slct_tbl_full_df01.sort_values(by=['source_name'])
#slct_tbl_full_df02 = slct_tbl_full_df02.reset_index()

print(f'DF instances: {len(slct_tbl_full_df02)}')

slct_tbl_full_df02['article_parsed'] = ''

total_urls = len(slct_tbl_full_df02)

# Start timer
start_time = dt.datetime.today()

# Use multiple parsing methods to extract the article from raw HTML
for i, row in enumerate(slct_tbl_full_df02.itertuples(), 1):
    #print(f'Enumeration #: {i}')
    #print(row[7])
    soup = BeautifulSoup(row[7], 'html.parser')

    # Check for available JSON object
    try:
        script_tag = soup.find('script', {'type': 'application/ld+json'})
        if script_tag == None:
            json_err01 = f'''JSON Object = None: Index: {i-1}; source: {row[2]};
            URL: {row[5]}'''
        else:
            article_json = json.loads(script_tag.string)
```

```

        article_content = article_json['articleBody']
        slct_tbl_full_df02.at[row.Index, 'article_parsed'] = article_content

# If no JSON object available, use BeautifulSoup to look for available
# HTML tags
except TypeError:
    print(f'Type Error')

except KeyError:
    json_err02 = f'''Missing JSON key: Index: {i-1}; source: {row[2]};
    URL: {row[5]}'''
    article_body = soup.find('div', class_='article__content-container')

    if article_body is None: #forfoxandbreitbart(sometimes)
        #print('Class != article__content-container')
        article_body = soup.find('p', class_="speakable")
        if article_body is None: #breitbart(most)
            #print('Class != speakable')
            article_body = soup.find('div', class_='entry-content')
            if article_body is None: #WashPost
                #print('Class != entry-content')
                article_body = soup.find('div', class_='article-body')

    if article_body is not None:
        article_text = article_body.get_text()
        slct_tbl_full_df02.at[row.Index, 'article_parsed'] = article_text
        #print('Rejoice, parse was successful!')
    else:
        print('\nParse not successful...')
        try:
            print(json_err02)
        except:
            pass

    print('.', end='')

# End timer script
end_time = dt.datetime.today()
time_elapse = end_time - start_time
print(f'Start Time = {start_time}')
print(f'End Time = {end_time}')
print(f'Elapsed Time = {time_elapse}')

```

DF instances: 6153

...

Parse not successful...

Missing JSON key: Index: 64; source: CNN;

URL: <https://www.cnn.com/politics/live-news/us-debt-ceiling-deadline->

talks-05-31-23/index.html

...

Parse not successful...

Missing JSON key: Index: 87; source: CNN;

URL: <https://www.cnn.com/politics/live-news/us-debt-ceiling-senate-vote-06-01-23/index.html>

...

...

...

...

Parse not successful...

Missing JSON key: Index: 407; source: CNN;

URL: <https://www.cnn.com/politics/live-news/us-debt-ceiling-deadline-talks-05-30-23/index.html>

...

...

...

...

Parse not successful...

Missing JSON key: Index: 665; source: CNN;

URL: <https://www.cnn.com/europe/live-news/russia-ukraine-war-news-05-06-23/index.html>

...

...

...

...

Parse not successful...

Missing JSON key: Index: 948; source: The Washington Post;

URL: <https://www.washingtonpost.com/dc-md-va/interactive/2023/proud-boys-trial-timeline-jan6-videos-chats/>

...

...

...

Parse not successful...

Missing JSON key: Index: 1122; source: CNN;

URL: <https://www.cnn.com/us/live-news/brownsville-texas-car-crash-05-08-23/index.html>

...

...

...

Parse not successful...

Missing JSON key: Index: 1341; source: CNN;

URL: <https://www.cnn.com/politics/live-news/george-santos-federal-charges-05-10-23/index.html>

...

...

...

...

...

...

...

...

...

...

Parse not successful...

Missing JSON key: Index: 2065; source: The Washington Post;
URL: <https://www.washingtonpost.com/nation/interactive/2023/texas-title-42-end/>

...

Parse not successful...

Missing JSON key: Index: 2101; source: CNN;
URL: <https://www.cnn.com/politics/live-news/trump-cnn-town-hall/index.html>

...

...

...

Parse not successful...

Missing JSON key: Index: 2290; source: CNN;
URL: <https://www.cnn.com/politics/live-news/trump-russia-probe-durham-report/index.html>

...

...

...

Parse not successful...

Missing JSON key: Index: 2513; source: CNN;
URL: <https://www.cnn.com/europe/live-news/russia-ukraine-war-news-05-16-23/index.html>

...

...

...

...

...

...

...

Parse not successful...

Missing JSON key: Index: 3040; source: CNN;
URL: <https://www.cnn.com/politics/live-news/biden-mccarthy-meeting-debt-ceiling-05-16-23/index.html>

...

...

...

Parse not successful...

Missing JSON key: Index: 3225; source: CNN;
URL: <https://www.cnn.com/europe/live-news/russia-ukraine-war-news-06-03-23/index.html>

...

...

...

...

...

Parse not successful...

Missing JSON key: Index: 3566; source: The Washington Post;
 URL: <https://www.washingtonpost.com/elections/candidates/christie-2024/>

...

Parse not successful...

Missing JSON key: Index: 3611; source: The Washington Post;
 URL: <https://www.washingtonpost.com/elections/candidates/mike-pence-2024/>

...

Parse not successful...

Missing JSON key: Index: 3650; source: CNN;
 URL: <https://www.cnn.com/politics/live-news/nikki-haley-cnn-town-hall/index.html>

...

...

...

Parse not successful...

Missing JSON key: Index: 3839; source: The Washington Post;
 URL: <https://www.washingtonpost.com/elections/candidates/doug-burgum-2024/>

...

...

...

Parse not successful...

Missing JSON key: Index: 4037; source: CNN;
 URL: <https://www.cnn.com/politics/live-news/mike-pence-cnn-town-hall/index.html>

...

...

...

...

Parse not successful...

Missing JSON key: Index: 4337; source: CNN;
 URL: <https://www.cnn.com/politics/live-news/mar-a-lago-documents-probe-latest/index.html>

...

Parse not successful...

Missing JSON key: Index: 4414; source: CNN;
 URL: <https://www.cnn.com/politics/live-news/trump-indictment-documents-06-11-23/index.html>

...

...

...

Parse not successful...
Missing JSON key: Index: 4622; source: CNN;
URL: <https://www.cnn.com/audio/podcasts/one-thing/episodes/3c097086-458f-47fb-baa7-b01d00348cce>
...
...
Parse not successful...
Missing JSON key: Index: 4753; source: The Washington Post;
URL: <https://www.washingtonpost.com/politics/interactive/2023/door-county-bellwether-politics/>
...
Parse not successful...
Missing JSON key: Index: 4812; source: CNN;
URL: <https://www.cnn.com/politics/live-news/trump-indictment-classified-documents-06-09-23/index.html>
...
...
...
Parse not successful...
Missing JSON key: Index: 5106; source: CNN;
URL: <https://www.cnn.com/politics/live-news/donald-trump-indictment-court-appearance-06-13-23/index.html>
...
...
Parse not successful...
Missing JSON key: Index: 5212; source: CNN;
URL: <https://www.cnn.com/politics/live-news/trump-indictment-documents-06-12-23/index.html>
...
...
Parse not successful...
Missing JSON key: Index: 5371; source: CNN;
URL: <https://www.cnn.com/politics/live-news/chris-christie-town-hall/index.html>
...
...
...
...
...
...
...
...
...
...
Parse not successful...
Missing JSON key: Index: 6114; source: Slate Magazine;
URL: <https://slate.com/news-and-politics/2023/06/neil-gorsuch-so-good->

native-americans-scotus.html

.

Parse not successful...

Missing JSON key: Index: 6115; source: Slate Magazine;

URL: <https://slate.com/news-and-politics/2023/06/the-slatest-june-sixteenth.html>

.

Parse not successful...

Missing JSON key: Index: 6116; source: Slate Magazine;

URL: <https://slate.com/news-and-politics/2023/06/amy-coney-barrett-supreme-court-native-win.html>

.

Parse not successful...

Missing JSON key: Index: 6117; source: Slate Magazine;

URL: <https://slate.com/news-and-politics/2023/06/mccarthy-house-republicans-freedom-caucus-budget-shutdown.html>

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6119; source: Slate Magazine;

URL: <https://slate.com/business/2023/06/cava-stock-ipo-investors-fast-casual-mediterranean.html>

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6121; source: Slate Magazine;

URL: <https://slate.com/news-and-politics/2023/06/daniel-ellsberg-dead-pentagon-papers-vietnam-war.html>

.

Parse not successful...

Missing JSON key: Index: 6122; source: Slate Magazine;

URL: <https://slate.com/news-and-politics/2023/06/iran-us-informal-nuclear-talks-what-it-means.html>

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6124; source: Slate Magazine;

URL: <https://slate.com/business/2023/06/francis-suarez-president-republicans-miami-crypto.html>

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6126; source: Slate Magazine;

URL: <https://slate.com/human-interest/2023/06/supreme-court-affirmative-action-decisions-race.html>

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6128; source: Slate Magazine;
 URL: <https://slate.com/culture/2023/06/cormac-mccarthy-dead-garbage-el-paso-texas.html>

.

Parse not successful...

Missing JSON key: Index: 6129; source: Slate Magazine;
 URL: <https://slate.com/culture/2023/06/black-mirror-season-6-joan-is-awful-netflix.html>

.

Parse not successful...

Missing JSON key: Index: 6130; source: Slate Magazine;
 URL: <https://slate.com/human-interest/2023/06/pride-protests-republicans-protests-muslims-michigan.html>

.Type Error

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6133; source: Slate Magazine;
 URL: <https://slate.com/podcasts/amicus/2023/06/justice-barretts-indian-child-welfare-act-indigenous-rights>

.Type Error

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6136; source: Slate Magazine;
 URL: <https://slate.com/news-and-politics/2023/06/missouri-anti-trans-legislation-judaism-testimony.html>

.Type Error

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6139; source: Slate Magazine;
 URL: <https://slate.com/news-and-politics/2023/06/the-slatest-news-and-politics-newsletter-catch-up-on-slates-top-stories-from-wednesday-june-15.html>

.Type Error

.Type Error

.

Parse not successful...

Missing JSON key: Index: 6142; source: Slate Magazine;
 URL: <https://slate.com/news-and-politics/2023/06/democrats-supreme-court-survey-voters-mad.html>

.

Parse not successful...

Missing JSON key: Index: 6143; source: Slate Magazine;
 URL: <https://slate.com/news-and-politics/2023/06/supreme-court-donald-trump-jack-smith-florida.html>

.Type Error

.Type Error

```

.
Parse not successful...
Missing JSON key: Index: 6146; source: Slate Magazine;
    URL: https://slate.com/news-and-politics/2023/06/trump-classified-
documents-case-indictment-dangers-questions.html
.Type Error
.
Parse not successful...
Missing JSON key: Index: 6148; source: Slate Magazine;
    URL: https://slate.com/news-and-politics/2023/06/target-starbucks-pride-
threats.html
.Type Error
.
Parse not successful...
Missing JSON key: Index: 6150; source: Slate Magazine;
    URL: https://slate.com/podcasts/political-gabfest/2023/06/aileen-cannon-
presides-in-a-donald-trump-case-again-political-gabfest
.Type Error
.
Parse not successful...
Missing JSON key: Index: 6152; source: Slate Magazine;
    URL: https://slate.com/podcasts/the-waves/2023/06/menstruation-how-
white-supremacy-and-the-patriarchy-ruined-a-very-normal-time-of-the-month
.Start Time = 2023-06-18 14:01:20.481350
End Time = 2023-06-18 14:08:36.000522
Elapsed Time = 0:07:15.519172

```

```

[7]: # Review sample
display(slct_tbl_full_df02.iloc[6119])
#display(slct_tbl_full_df02.iloc[6119]['article_text'])

```

```

text_id                                8413
source_name                            Slate Magazine
author                                Nitish Pahwa
title      Wall Street's Newest Darling Is an Unprofitabl...
url      https://slate.com/business/2023/06/cava-stock-...
publish_date      2023-06-16T22:22:57Z
article_text      <!DOCTYPE html>\n<html data-env="prod" data-la...
content      Stock markets aint what they used to be. Since...
article_parsed
Name: 6119, dtype: object

```

```

[8]: display(slct_tbl_full_df02[slct_tbl_full_df02['article_parsed']==''].head(11))
print(len(slct_tbl_full_df02[slct_tbl_full_df02['article_parsed']=='']))

```

```

      text_id      source_name \
64      65      CNN
74      75      Fox News
87      88      CNN

```

255	257	CNN
407	566	CNN
541	760	Fox News
665	1173	CNN
948	1456	The Washington Post
1122	1737	CNN
1341	2025	CNN
2065	2749	The Washington Post

	author \
64	By Adrienne ...
74	Sofie Watson
87	By Maure...
255	NaN
407	By Mike Hayes and <a href="/profiles/maureen-c...
541	Danielle Tuntigian
665	By Sophie Tanno
948	Adriana Usero
1122	By Aditi Sang...
1341	By Adrienne ...
2065	Arelis Hernández, Whitney Shefte, Jabin Botsford

	title \
64	The latest on the US debt ceiling deal
74	FOX NEWS CHANNEL TO DEBUT NEW WEEKEND PRIMETIM...
87	The latest on the US debt ceiling deal
255	UPDATE: WAR ON GAY PRIDE...
407	The latest on the US debt ceiling deal
541	SEN. JONI ERNST AND SEN. JEANNE SHAHEEN TO PAR...
665	Russia's war in Ukraine
948	Proud Boys revealed: Videos, secret chats show...
1122	8 killed when driver plows into crowd outside ...
1341	George Santos charged in federal probe
2065	Texas uses aggressive tactics to arrest...

	url	publish_date \
64	https://www.cnn.com/politics/live-news/us-debt...	2023-05-31T12:35:50Z
74	https://press.foxnews.com/2023/06/fox-news-cha...	2023-06-01T18:33:43Z
87	https://www.cnn.com/politics/live-news/us-debt...	2023-06-01T13:46:13Z
255	https://view.newsletters.cnn.com/messages/1685...	2023-06-01T21:55:37Z
407	https://www.cnn.com/politics/live-news/us-debt...	2023-05-30T14:12:09Z
541	https://press.foxnews.com/2023/05/sen-joni-ern...	2023-05-30T17:07:24Z
665	https://www.cnn.com/europe/live-news/russia-uk...	2023-05-06T09:46:06Z
948	https://www.washingtonpost.com/dc-md-va/intera...	2023-05-05T16:00:23Z
1122	https://www.cnn.com/us/live-news/brownsville-t...	2023-05-08T14:28:31Z
1341	https://www.cnn.com/politics/live-news/george-...	2023-05-10T12:52:24Z
2065	https://www.washingtonpost.com/nation/interact...	2023-05-10T23:25:36Z

