# Web scraping using Requests-HTML

Github link: <https://github.com/aditiabhang/my-python-stuff/tree/master/my-projects/tutorial_project>

**Introduction:** This program code about how to scrape websites using the Requests-HTML library.

Requests-HTML is an excellent tool for parsing HTML code and grabbing exactly the information you need. So whether you're pulling down headlines from news sites, scores from sports websites, or prices from an online store

import csv # to write the output into a csv file

from requests\_html import HTML, HTMLSession

csv\_file = open('cms\_scaper.csv', 'w')

csv\_writer = csv.writer(csv\_file)

csv\_writer.writerow(['Headline', 'Summary', 'Video'])

session = HTMLSession()

response\_obj = session.get('https://coreyms.com/')

#print(response\_obj.html)

articles = response\_obj.html.find('article')

for article in articles:

# print(article.html) # prints out html of the first article

# adding the for loop, because we want all the articles now..and removing first = True

headline = article.find('.entry-title-link', first = True).text

print('\nHeadline: ', headline)

print("")

summary = article.find('.entry-content p', first = True).text

print('Summary: ', summary)

print("")

#video\_src = article.find('iframe', first = True)

#print(video\_src.attrs['src']) #attrs = attributes, just like dictionary ,

# and the 'src' is a key

#print(video\_src)

# try except block to check for errors/missing attributes

try:

video\_src = article.find('iframe', first = True).attrs['src']

# Now we get the ID present in the url by using split method

# The ID is present at the 4th index, hence use [4]

video\_id = video\_src.split('/')[4]

video\_id = video\_id.split('?')[0]

#print('YOUTUBE Video ID: ',video\_id)

yt\_link = f'https://www.youtube.com/watch?v={video\_id}'

except Exception as e:

t\_link = None

print('YOUTUBE link: ', yt\_link)

print("")

csv\_writer.writerow([headline, summary, yt\_link])

csv\_file.close()

'''

Now we loop over all of the articles and get that information for all of them.

We add the for loop.

Now sometimes, there can be a case when the find method cannot find a attribute in the list, in such case we try except block.

This is check for errors, if found it will just skip that part.

'''