scraping

June 11, 2024

This was all about data scrapping from a lawyer website, where I was to get personal data (name. company,phone,website,and location). This was an easy part as it involved using selenium together with chrome driver, then I used beautiful-soup to find the elements I needed in the DOM. A tricky part came in where I was to process the data, the dynamic rendering of the web made it a bit tricky but I managed in cleaning the data. First it was by finding the links to their profile, the visiting their profile to find data in a single request. All was left was saving the data in a CSV which I did with ease with the pandas framework.

[ ]:
```
# importing the needed libraries
from selenium import webdriver
from bs4 import BeautifulSoup as bs
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
import time
import random
import pandas as pd
```

[ ]:
```
# Set up Chrome options
chrome_options = Options()
chrome_options.add_argument("--headless")

driver = webdriver.Chrome()
```

[ ]:
```
x,y = 1,1

l_ = open("links.txt","a")
df = pd.DataFrame(columns=["name", "company", "phone", "link", "address"])

while True:
# function to return source code
    url = f"https://www.avvo.com/personal-injury-lawyer/ca.html?page={x}"
    print(url)
    def next_page(x):
        x +=1
        driver.get(url)

        # Give the page some time to load
```

```
        wait = driver.implicitly_wait(20)
        page_source = driver.page_source

        if not page_source:
            print("its the END!")
        return page_source, x

    page_source , x = next_page(x)

    # Open a file in write mode
    with open("page_source.html", "w", encoding="utf-8") as f:
        f.write(page_source)

    # # Close the WebDriver
    # driver.quit()

    # Parse the HTML source using BeautifulSoup
    soup = bs(page_source, "html.parser")

    # Find all ad elements
    ads = soup.find_all(class_="gtm-tracking-container")
    link_ = []


    for ad in ads:
        links = ad.find_all("a")
        for i in range(len(links)):
            link = links[i]["href"]
            link_.append(link)
            driver.get(link)
            source = driver.page_source
            try:
                s = bs(source,"html.parser")
                name = s.find(class_="lawyer-name").text.strip()
                company = s.find(class_="contact-firm").text.strip()
                phone = s.find(class_="overridable-lawyer-phone-copy").text.
↪strip()
                # fax = s.find(class_="phone_call_initiated")
                website = s.find(class_="contact-website")
                _addresses = s.find(class_="contact-address").find("div")
                address = "".join(_ad.text.strip().replace("#",",") for _ad in␣
↪_addresses)
            except AttributeError:
                pass

            try:
```

```
                a = website.find("a")["onclick"].strip("this.href=").strip(";
 ↪return true;")
                # comb_ = f"name:{name},company:{company},phone:{phone},link:
 ↪{a} ,address:{address}"
                comb_ = {"name":name,"company":company,"phone":phone,"link":
 ↪a,"address":address}

            except AttributeError as e:
                comb_ = {"name":name,"company":company,"phone":phone,"address":
 ↪address}

            df = df._append(comb_,ignore_index=True)
            # l_.write(comb_ + "\n")
            break
    df.to_csv("data.csv", index="name")

    time.sleep(random.uniform(3, 8))
# print(link_)
```

0.1 On the terminal it should print out the page links

https://www.avvo.com/personal-injury-lawyer/ca.html?page=1 https://www.avvo.com/personal-injury-lawyer/ca.html?page=2 https://www.avvo.com/personal-injury-lawyer/ca.html?page=3 https://www.avvo.com/personal-injury-lawyer/ca.html?page=4 https://www.avvo.com/personal-injury-lawyer/ca.html?page=5 https://www.avvo.com/personal-injury-lawyer/ca.html?page=6 https://www.avvo.com/personal-injury-lawyer/ca.html?page=7 https://www.avvo.com/personal-injury-lawyer/ca.html?page=8 https://www.avvo.com/personal-injury-lawyer/ca.html?page=9 https://www.avvo.com/personal-injury-lawyer/ca.html?page=10 https://www.avvo.com/personal-injury-lawyer/ca.html?page=11 https://www.avvo.com/personal-injury-lawyer/ca.html?page=12 https://www.avvo.com/personal-injury-lawyer/ca.html?page=13 https://www.avvo.com/personal-injury-lawyer/ca.html?page=14 https://www.avvo.com/personal-injury-lawyer/ca.html?page=15 https://www.avvo.com/personal-injury-lawyer/ca.html?page=16 https://www.avvo.com/personal-injury-lawyer/ca.html?page=17 https://www.avvo.com/personal-injury-lawyer/ca.html?page=18 https://www.avvo.com/personal-injury-lawyer/ca.html?page=19 https://www.avvo.com/personal-injury-lawyer/ca.html?page=20 https://www.avvo.com/personal-injury-lawyer/ca.html?page=21 https://www.avvo.com/personal-injury-lawyer/ca.html?page=22 https://www.avvo.com/personal-injury-lawyer/ca.html?page=23 https://www.avvo.com/personal-injury-lawyer/ca.html?page=24 https://www.avvo.com/personal-injury-lawyer/ca.html?page=25

0.2 This is the data will be saved on the data.csv file

0,Michael Joseph Cefali,Cefali & Cefali,(855) 925-1267,‘https://www.avvo.com/attorney-website/92675-ca-michael-cefali-4872332/website.html’,“27130 Paseo EspadaBldg B Suite , 521San Juan Capistrano, CA, 92675” 1,Edgar Poghosyan,Attorneys Incorporated,(877) 797-0048,„“535 N Brand Boulevard, 5th FlGlendale, CA, 91203” 2,Christopher B. Adamson,Adamson Ahdoot

LLP,(855) 808-5010,‘https://www.avvo.com/attorney-website/90035-ca-christopher-adamson-93218/website.html’,“3165 Olin Ave,2San Jose, CA, 95117” 3,Kevin Crockett,Text Kevin Accident Attorneys,(714) 714-7100,‘https://www.avvo.com/attorney-website/92618-ca-kevin-crockett-4872506/website.html’,“7700 Irvine Center DrSte 400Irvine, CA, 92618” 4,Arsen Sarapinian,“Law Offices of Arsen Sarapinian, P.C.”,(213) 348-1578,‘https://www.avvo.com/attorney-website/90212-ca-arsen-sarapinian-4574853/website.html’,“9350 Wilshire Blvd., Ste. 203Beverly Hills, CA, 90212” 5,Alan A. Ahdoot,Adamson Ahdoot LLP,(844) 982-2537,‘https://www.avvo.com/attorney-website/90035-ca-alan-ahdoot-47251/website.html’,“3165 Olin Ave,2San Jose, CA, 95117” 6,S. David Rosenthal,Rosenthal Law,(916) 459-2003,‘https://www.avvo.com/attorney-website/95811-ca-s-rosenthal-372222/website.html’,“2251 Douglas Blvd.,120Roseville, CA, 95661” 7,Haleh Shekarchian,Law Offices of Haleh Shekarchian,(310) 789-2192,‘https://www.avvo.com/attorney-website/90210-ca-haleh-shekarchian-349005/website.html’,“9440 Santa Monica Boulevard Suite 707Beverly Hills, CA, 90210” 8,Federico C. Sayre,Adamson Ahdoot LLP,(844) 729-1091,‘https://www.avvo.com/attorney-website/92618-ca-federico-sayre-269768/website.html’,“8895 Research DriveSuite 100Irvine, CA, 92618” 9,Emery Brett Ledger,The Ledger Law Firm | Accident Attorneys,(866) 249-1408,‘https://www.avvo.com/attorney-website/92660-ca-emery-ledger-252531/website.html’,“5160 Birch StreetSuite 100Newport Beach, CA, 92660” 10,Joshua Michael Bonnici,“BONNICI LAW GROUP, APC”,(619) 259-5460,‘https://www.avvo.com/attorney-website/92101-ca-joshua-bonnici-4069370/website.html’,“1620 5th AvenueSuite 625San Diego, CA, 92101” 11,Austin G. Ward,Adamson Ahdoot LLP,(855) 337-6161,‘https://www.avvo.com/attorney-website/90035-ca-austin-ward-4645970/website.html’,“1150 S Robertson BoulevardLos Angeles, CA, 90035” 12,John Gary Jahrmarkt,Jahrmarkt & Associates,(310) 907-8450,‘https://www.avvo.com/attorney-website/90067-ca-john-jahrmarkt-117897/website.html’,“2049 Century Park E Ste 3850Los Angeles, CA, 90067” 13,James Otto Heiting,Heiting & Irwin,(951) 221-8286,‘https://www.avvo.com/attorney-website/92506-ca-james-heiting-236838/website.html’,“6216 Brockton AveSuite 111Riverside, CA, 92506” 14,Roger S. Bonakdar,Bonakdar Law Firm,(559) 495-9233,‘https://www.avvo.com/attorney-website/93721-ca-roger-bonakdar-1753422/website.html’,“2344 Tulare St., Suite 301Fresno, CA, 93721-2242”