Web Crawling BAB

June 13, 2024

0.1 Importing necessary packages

[1]:

```
from requests_html import HTMLSession
from bs4 import BeautifulSoup
import pandas as pd
import time
import requests
```

0.2 Starting HTMLSession and BeautifulSoup and ensuring both are working

[5]:

```
url = 'https://muqawil.org/en/contractors'
session = HTMLSession()
#response = requests.get(url) #for BeautifulSoup
r = session.get(url) # for HTMLSession
r.html.arender(sleep=1)
soup = BeautifulSoup(r.html.html, 'html.parser')
print(r.status_code)
#print(response.status_code)
```

```
200
```

```
C:\Users\s2ahm\AppData\Local\Temp\ipykernel_14300\986819458.py:5:
RuntimeWarning: coroutine 'HTML.arender' was never awaited
  r.html.arender(sleep=1)
RuntimeWarning: Enable tracemalloc to get the object allocation traceback
```

0.3 Finding the xpath for the container of all contractors

[6]:

```
companies = r.html.xpath('//*[@id="all_contractor"]', first=True)
print(companies)
```

```
<Element 'div' class=('col-lg-8', 'col-xl-9') id='all_contractor'>
```

0.4 Ensuring the links of contractors are clear and valid

[7]:

```
print(companies.absolute_links)
```

```
{'https://muqawil.org/en/contractors/20006199/143',
'https://muqawil.org/en/contractors/959/143',
'https://muqawil.org/en/contractors/20019229/143',
 'https://muqawil.org/en/contractors/20005421/143',
 'https://muqawil.org/en/contractors/20010655/143',
 'https://muqawil.org/en/contractors/20016964/143',
 'https://muqawil.org/en/contractors?page=3',
 'https://muqawil.org/en/contractors/20004514/143',
 'https://muqawil.org/en/contractors/20001440/143',
 'https://muqawil.org/en/contractors/20020449/143',
 'https://muqawil.org/en/contractors/869/143',
 'https://muqawil.org/en/contractors?page=903',
 'https://muqawil.org/en/contractors/20010987/143',
 'https://muqawil.org/en/contractors/8050/143',
 'https://muqawil.org/en/contractors?page=1',
 'https://muqawil.org/en/contractors/20003088/143',
 'https://muqawil.org/en/contractors/20002330/143',
 'https://muqawil.org/en/contractors/20008800/143',
 'https://muqawil.org/en/contractors/8649/143',
 'https://muqawil.org/en/contractors/20008518/143',
 'https://muqawil.org/en/contractors/20012033/143',
 'https://muqawil.org/en/contractors?page=2',
 'https://muqawil.org/en/contractors/20008004/143',
 'https://muqawil.org/en/contractors/20023122/143'}
```

[8]:

```python
for item in companies.absolute_links:
    print(item)
```

```
https://muqawil.org/en/contractors/20006199/143
https://muqawil.org/en/contractors/959/143
https://muqawil.org/en/contractors/20019229/143
https://muqawil.org/en/contractors/20005421/143
https://muqawil.org/en/contractors/20010655/143
https://muqawil.org/en/contractors/20016964/143
https://muqawil.org/en/contractors?page=3
https://muqawil.org/en/contractors/20004514/143
https://muqawil.org/en/contractors/20001440/143
https://muqawil.org/en/contractors/20020449/143
https://muqawil.org/en/contractors/869/143
https://muqawil.org/en/contractors?page=903
https://muqawil.org/en/contractors/20010987/143
https://muqawil.org/en/contractors/8050/143
https://muqawil.org/en/contractors?page=1
https://muqawil.org/en/contractors/20003088/143
https://muqawil.org/en/contractors/20002330/143
https://muqawil.org/en/contractors/20008800/143
https://muqawil.org/en/contractors/8649/143
https://muqawil.org/en/contractors/20008518/143
https://muqawil.org/en/contractors/20012033/143
https://muqawil.org/en/contractors?page=2
https://muqawil.org/en/contractors/20008004/143
```

https://muqawil.org/en/contractors/20023122/143

0.5 Decoding the emails as they are protected by the website

[9]:
```
def cfDecodeEmail(encodedString):
    '''
    Emails are protected in the website
    this function decodes the emails
    '''
    r = int(encodedString[:2],16)
    email = ''.join([chr(int(encodedString[i:i+2], 16) ^ r) for i in range(2,␣
↪len(encodedString), 2)])
    return email
```

0.6 Creating scrape_company function to scrape the required data and extract it to a pandas dataframe and an excel file

note: running this cell will take +5 minutes to finish

[12]:
```
def scrape_company(item, page_number, data):
    '''''
    item represents the href or url for each company
    page_number will be created later
    data is the variables which need to be scraped like name, phone, email..etc
    '''''
    r = session.get(item)
    r.html.arender(sleep=1)
    soup = BeautifulSoup(r.html.html, 'html.parser')

    # node_n for company name
    # node_c for city
    # node_e for email
    # node_p for phone
    # activities_nodes for activities

    node_n = r.html.xpath('/html/body/main/div/div/div/div/div/div[2]/h3',␣
↪first=True)
    company_name = node_n.text if node_n is not None else None
    node_c = soup.find('div', attrs={'class': 'info-name'}, string='City')
    city = node_c.find_next('div', class_='info-value').text.strip() if node_c␣
↪is not None else None
    node_e = r.html.xpath('/html/body/main/div/div/div/div/div/div[2]/div[2]/
↪div/div[7]/div/div[2]/div[2]/a', first=True)
    email = cfDecodeEmail(node_e.attrs['href'].replace('/cdn-cgi/l/
↪email-protection#', '')) if node_e is not None else None
    node_p = soup.find('div', attrs={'class': 'info-name'}, string='phone')
    phone = node_c.find_next('div', class_='info-value').text.strip() if node_c␣
 ↪is not None else None
    activities_nodes = soup.find_all('li', class_='list-item')
    activities = ', '.join([node.text.strip() for node in activities_nodes]) if␣
 ↪activities_nodes else None

    if company_name and (city or email or phone or activities): # to avoid␣
 ↪getting None and null values
        data.append({
            'Page': page_number,
            'Company Name': company_name,
            'City': city,
            'Email': email,
            'Phone': phone,
            'Activities': activities
    })

# Initializing an empty list to collect data
data = []

base_url = 'https://muqawil.org/en/contractors'

# Iterating over the first 10 pages
for page in range(1, 11):
    page_url = f"{base_url}?page={page}"
    r = session.get(page_url)
    r.html.arender(sleep=1)

    companies = r.html.xpath('//*[@id="all_contractor"]', first=True)
    if not companies:
        continue

    for item in companies.absolute_links: # running scrape_company for each␣
 ↪company link in companies, takes +5 minutes
        scrape_company(item, page, data)

# Converting the data to a Pandas DataFrame
df = pd.DataFrame(data)

# Converting the DataFrame to an xlsx file and saving to locally
df.to_excel('contractors.xlsx', index=False)

# Displaying the first 40 rows of the DataFrame
df.head(40)
```

```
C:\Users\s2ahm\AppData\Local\Temp\ipykernel_14300\2920267594.py:47:
RuntimeWarning: coroutine 'HTML.arender' was never awaited
```

```
  r.html.arender(sleep=1)
RuntimeWarning: Enable tracemalloc to get the object allocation traceback
C:\Users\s2ahm\AppData\Local\Temp\ipykernel_14300\2920267594.py:8:
RuntimeWarning: coroutine 'HTML.arender' was never awaited
  r.html.arender(sleep=1)
RuntimeWarning: Enable tracemalloc to get the object allocation traceback
```

[12]:

```
     Page                                       Company Name  \
0       1                 Sharjah Development Contracting Co
1       1                      Alenjazat Contracting Company
2       1                      Dome Park Contracting Company
3       1                      acn solutions for contracting
4       1                   Bunoon Wa Funoon Contracting Co.
5       1                      On Al-Arabia Contracting Est.
6       1      Modern Building Solutions Contracting Company
7       1                 Ratel Al Sharq Contracting Company
8       1                       Almegren International Group
9       1             Gulf Pioneers for Construction Company
10      1             Manal Abdullah Al Anzi Contracting Est
11      1        Al-Manarat Al-Dhahabiya Contracting Company
12      1                    Rawasi Sama Contracting Company
13      1                               Tala Trading Company
14      1           Rasm Al Benaa Trading and Contracting Co
15      1  Atlas Cities General Architectural Contracting Co
16      1                 Awared General Contracting Company
17      1                        Expert Gate Contracting Est
18      1                White Beach General Contracting Est
19      1                      Gulf Pioneers Trading Company
20      2  Al-Asas Engineering Corporation for General Co…
21      2  Maha Ghadeer Ramadan Al Shammari General Contr…
22      2                       Access point for contracting
23      2                      Fineun Alamar Contracting Est
24      2                     My Best Choice Contracting Est
25      2                     Mosanadat Alemdad Company Ltd.
26      2  Mona Nazzal Dehaidah Al Harbi General Contract…
27      2  Mastery and Quality Architectural Contracting Est
28      2             New Lover Home General Contracting Est
29      2  Unique Execution Corporation for General Contr…
30      2                     Asas Al Taawon Contracting Est
31      2       Munira Nasser Omair Al Omair Contracting Est
32      2            Wissam Al-Washm General Contracting Est
33      2              Raissy Trading & Contracting Co. Ltd.
34      2  Future Houses General Architectural Contractin…
35      2    Asifat Al-Orouba Operation and Maintenance Est.
36      2                   Real estate direct broker office
37      2          Manarat AlOmran AlHadeeth Contracting Est.
38      2  Najmat Al Mimar General Contracting Company (o…
39     2                      Wasl Al Khair Contracting Est

                        City                          Email  \
0                  AL KHOBAR          alsharqimna@gmail.com
1                     RIYADH              info@alenjazat.sa
2         AL MUWAYH AL JADID           vv.com838@icloud.com
3                     JEDDAH            alwa7ed@hotmail.com
4                     RIYADH               info@bfconst.com
5                     RIYADH          a1032500371@gmail.com
6                     RIYADH         ttalshammari@gmail.com
7                     RIYADH            adel_77@hotmail.com
8                     RIYADH               info@almegren.sa
9                     RIYADH                   hh@gpksa.com
10                   AT TAIF       abdelaziz209@hotmail.com
11                   MUHAYIL       Moodel121212@hotmail.com
12            HAFAR AL BATIN                Ce3@hotmail.com
13                      HAIL          talaxtala@hotmail.com
14                    DAMMAM              amz@rasmco.com.sa
15            KHAMIS MUSHAYT          citiesatlas@gmail.com
16                    RIYADH            fared@fared-est.com
17                     TABUK        alhwaiti654@hotmail.com
18                    SAMTAH             mohd2536@gmail.com
19                    RIYADH                  acc@gpksa.com
20  AL MADINAH AL MUNAWWARAH             b2b2020@outlook.sa
21                  AL HUFUF          abda1zaal@hotmail.com
22                 AL KHAFJI     taha0595742980@outlook.com
23                    RABIGH        saeedstar88@hotmail.com
24                    RIYADH            Salnofaiy@gmail.com
25                              meshal@alkathiriholding.com
26                  BURAYDAH            f15-f20@hotmail.com
27                    JEDDAH  malki.abdulrahman80@gmail.com
28                  BURAYDAH              aljnshz@gmail.com
29                     JAZAN          qanai1234@hotmail.com
30                    RIYADH          hmd_55111@hotmail.com
31            HAFAR AL BATIN      aazzsaleh882364@gmail.com
32                   SHAQRA'          bader202005@gmail.com
33                    RIYADH            AHMAD@RAISSY.COM.SA
34                    RIYADH           f.h.c.9282@gmail.com
35                   MUHAYIL      capital.arabism@gmail.com
36                      HAIL             hawas770@gmail.com
37            KHAMIS MUSHAYT          Alshehry300@gmail.com
38                    RIYADH            nasy-11@outlook.com
39         AHAD AL MUSARIHAH        habcdefgh2012@gmail.com

                       Phone  \
0                  AL KHOBAR
1                     RIYADH
```

```
2         AL MUWAYH AL JADID
3                     JEDDAH
4                     RIYADH
5                     RIYADH
6                     RIYADH
7                     RIYADH
8                     RIYADH
9                     RIYADH
10                   AT TAIF
11                   MUHAYIL
12            HAFAR AL BATIN
13                      HAIL
14                    DAMMAM
15            KHAMIS MUSHAYT
16                    RIYADH
17                     TABUK
18                    SAMTAH
19                    RIYADH
20  AL MADINAH AL MUNAWWARAH
21                  AL HUFUF
22                 AL KHAFJI
23                    RABIGH
24                    RIYADH
25                          
26                  BURAYDAH
27                    JEDDAH
28                  BURAYDAH
29                     JAZAN
30                    RIYADH
31            HAFAR AL BATIN
32                   SHAQRA'
33                    RIYADH
34                    RIYADH
35                   MUHAYIL
36                      HAIL
37            KHAMIS MUSHAYT
38                    RIYADH
39         AHAD AL MUSARIHAH

                                           Activities
0   Construction of buildings, Construction of bui…  
1   Construction of buildings, Construction of bui…  
2   Construction of buildings, Construction of bui…  
3   Waste collection, treatment & disposal activit…  
4   Construction of buildings, Construction of bui…  
5                                             No Data
6   Construction of buildings, Construction of bui…  
```

```
7   Construction of buildings, Construction of bui…
8   Construction of buildings, Construction of bui…
9   Construction of buildings, Construction of bui…
10  Construction of buildings, Construction of bui…
11                                              No Data
12  Construction of buildings, Construction of bui…
13  Construction of buildings, Construction of bui…
14  Waste collection, treatment & disposal activit…
15  Construction of buildings, Construction of bui…
16  Construction of buildings, Construction of bui…
17  Mining support services, Oil and natural gas e…
18  Construction of buildings, Construction of bui…
19  Construction of buildings, Construction of bui…
20                                              No Data
21  Construction of buildings, Construction of bui…
22                                              No Data
23                                              No Data
24  Waste collection, treatment & disposal activit…
25  Construction of buildings, Construction of bui…
26  Construction of buildings, Construction of bui…
27  Construction of buildings, Construction of bui…
28  Construction of buildings, Construction of bui…
29  Construction of buildings, Construction of bui…
30                                              No Data
31  Construction of buildings, Construction of bui…
32  Construction of buildings, Construction of bui…
33  Construction of buildings, Construction of bui…
34  Construction of buildings, Construction of bui…
35  Construction of buildings, Construction of bui…
36  Mining support services, Oil and natural gas e…
37  Construction of buildings, Construction of bui…
38  Construction of buildings, Construction of bui…
39  Construction of buildings, Construction of bui…
```

0.6.1 Challenges

1. Time:

   As my first project in web scraping, it took me much time to finish the web crawling process, I couldn't do the semantic search challenge because of lack of time. * * *

2. Dynamic javascript for email scraping:

   I also did not have the time to learn selenium to scrape emails, so instead I used a decoding function.

Thanks for reading * * by Ahmed Sharabati*