1canpmfpt

January 29, 2024

Web scrapping

En este proyecto se extrae información de un sitio argentino de venta de casas (https://www.argenprop.com/casa-venta-localidad-capital-federal) usando la libreria de Python Beautiful Soup. A continuación se detalla el proceso para realizar el scrapping:

Conexión a la página con request

Creación de la sopa y extracción de información

Tratamiento de las columnas y creación del dataframe

Extracción de información del total de páginas del sitio

[15]:
```
import pandas as pd
import numpy as np
import requests
from bs4 import BeautifulSoup
from lxml import etree
from datetime import date
from datetime import datetime
import seaborn as sns
```

1. Conexión a la página con request

[16]:
```
url = requests.get('https://www.argenprop.com/
 ↪casa-venta-localidad-capital-federal')
url.status_code
```

[16]: 200

2. Creación de la sopa y extracción de información

[3]:
```
#Se extrae la información del html

casas = BeautifulSoup(url.content, 'html.parser')
dom= etree.HTML(str(casas))

#Se extrae la ubicación

ubicacion = casas.find_all('p', attrs={"class":"card__address"})
```

```
ubicacion=[i.text for i in ubicacion]

#Se extrae la descripción base

descripcion_base = casas.find_all('h2', attrs={"class":"card__title"})
descripcion_base=[i.text for i in descripcion_base]

#Se extrae el área

area=dom.xpath('//ul[@class="card__main-features"]/li[1]/span')
area=[i.text for i in area]

#Se extraen los dormitorios

dormitorios=dom.xpath('//ul[@class="card__main-features"]/li[2]/span')
dormitorios=[i.text for i in dormitorios]

#Se extrae la antiguedad

antiguedad=dom.xpath('//ul[@class="card__main-features"]/li[3]/span')
antiguedad=[i.text for i in antiguedad]

#Se extrae el precio

precio = casas.find_all('p', attrs={'class':'card__price'})
precio=[i.text for i in precio]

#Se extrae la url

urls = casas.find_all('a', attrs={'class':'card'})
urls = [i.get('href') for i in urls]
urls=['https://www.argenprop.com'+i for i in urls]
```

3. Tratamiento de las columnas y creación del dataframe

[4]:
```
#Se crea el dataframe y se tratan las columnas correspondientes

df = pd.DataFrame({'Descripción':descripcion_base,'Ubicación':ubicacion, 'Área':
 ↪area, 'Dormitorios':dormitorios,'Antigüedad':antiguedad, 'Precio':precio,␣
 ↪'Link':urls})
df.Precio=df.Precio.str.replace('\n','').str.replace(' ','').str.
 ↪replace('USD','USD  ').str.replace('.',',')
df.Ubicación=df.Ubicación.str.replace('\n','').str.strip(' ').str.strip('\n')
df.Área=df.Área.str.replace('\n','').str.strip(' ').str.strip('\n')
df.Antigüedad=df.Antigüedad.str.replace('\n','').str.strip(' ').str.strip('\n')
df.Dormitorios=df.Dormitorios.str.replace('\n','').str.strip(' ').str.
 ↪strip('\n')
```

```
df['Fecha_obs']=date.today()
```

[5]:
```
df
```

[5]:

	Descripción	Ubicación	Área
0	Casa 4 Ambientes Patio/ Galería/ Quincho sobre…	Pareja al 3800	200 m² cubie.
1	Impecable casa de 4 ambientes en 3 plantas - G…	Argerich 3717	185 m² cubie.
2	Casa - Villa Devoto	Mercedes al 4900	300 m² cubie.
3	Venta de Casa 5 ambientes con oficina y deposi…	Benito Juárez al 2900	330 m² cubie.
4	Venta - Casa 6 ambientes sobre lote propio de …	Gualeguaychu al 3014	164 m² cubie.
5	Venta - Casa 4 Ambientes en 2 Plantas con Gara…	Benito Juarez al 3400	156 m² cubie.
6	REGIA CASA . AMP LIVING COMEDOR. 2 DORMITORIOS…	J Navarro 4700	140 m² cubie.
7	Casa - en venta, Gran potencial! a Mts de Jon…	LOPEZ DE VEGA 1500	132 m² cubie.
8	Casa - en venta, Gran potencial! a Mts de Jon…	NEMESIO TREJO 5100	132 m² cubie.
9	CASA VILLA DEL PARQUE 4 AMB COCHERA Y PARRILLA.	Camarones 3000	142 m² cubie.
10	VENTA DUEÑO DIRECTO - CASA LOTE PROPIO	Pje Calingasta 1800	175 m² cubie.
11	Triplex 4 Ambientes - Villa Devoto	TRIPLEX 4 amb/2 COCHERAS- SIN EXPENSAS-PARRILL…	158 m² cubie.
12	CASA DE ESTILO 5 AMB. BCON.GARAJE, PATIO,PISCINA.	Emilio Lamarca 4900	291 m² cubie.
13	CASA LOTE PROPIO 4 DORMITOR-PATIO-COCHERA-PARR…	Bazurco 3300	143 m² cubie.
14	Casa - Venta - Argentina, Capital Federal - PI…	PICO 4900	240 m² cubie.
15	Casa Chalet en Venta ubicado en Saavedra, Cap…	Pinto 4600	146 m² cubie.
16	Casa en lote propio todo en bajos 3 amb	Fray G Arregui 3800	75 m² cubie.
17	Casa - en venta, Gran potencial! a Mts de Jon…	LOPEZ DE VEGA 1500	132 m² cubie.
18	Espectacular casa! Lote propio! 365 Mts! T/luz…	V Aguilar 2000	365 m² cubie.
19	CASAS - CASA - BELGRANO R, CAPITAL FEDERAL	SUCRE ANTONIO J DE MCAL. 3400	300 m² cubie.

	Dormitorios	Antigüedad	Precio
0	4 dorm.	18 años	USD 690,000
1	3 dorm.	12 años	USD 370,000
2	4 dorm.	8 años	USD 595,000
3	4 dorm.	15 años	USD 430,000
4	4 dorm.	20 años	USD 350,000
5	3 dorm.	30 años	USD 245,000
6	2 dorm.	15 años	USD 235,000
7	3 dorm.	90 años	USD 158,000
8	3 dorm.	90 años	USD 158,000
9	3 dorm.	22 años	USD 195,000
10	3 dorm.	1 año	USD 329,000
11	3 dorm.	A Estrenar	USD 495,000
12	3 dorm.	40 años	USD 315,000
13	4 dorm.	16 años	USD 200,000
14	5 dorm.	60 años	USD 380,000
15	3 dorm.	25 años	USD 380,000
16	2 dorm.	50 años	USD 165,000
17	3 dorm.	90 años	USD 158,000
18	5 dorm.	30 años	USD 460,000
19	3 dorm.	60 años	USD 690,000

	Link	Fecha_obs
0	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
1	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
2	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
3	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
4	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
5	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
6	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
7	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
8	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
9	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
10	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
11	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
12	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
13	https://www.argenprop.com/casa-en-venta-en-vil…	2024-01-29
14	https://www.argenprop.com/casa-en-venta-en-saa…	2024-01-29
15	https://www.argenprop.com/casa-en-venta-en-saa…	2024-01-29
16	https://www.argenprop.com/casa-en-venta-en-flo…	2024-01-29
17	https://www.argenprop.com/casa-en-venta-en-mon…	2024-01-29
18	https://www.argenprop.com/casa-en-venta-en-bel…	2024-01-29
19	https://www.argenprop.com/casa-en-venta-en-bel…	2024-01-29

4. Extracción de información del total de páginas del sitio

[6]:
```
#siguiente='https://www.argenprop.com'+dom.xpath('//ul[@class="pagination␣
 ↪pagination--links"]/li[contains(@class,"-next")]/a')[0].get('href')
```

[7]:
```
#ini=dom.xpath('//li[@class="pagination__page pagination__page--current"]/
 ↪span')[0].text
#ini=int(ini)
```

[8]:
```
#can=dom.xpath('//li[@class="pagination__page"][3]/a')[0].text
#can=int(can)
```

[42]:
```
l_descripcion=[]
l_ubicacion=[]
l_area=[]
l_dormitorios=[]
l_antiguedad=[]
l_precio=[]
l_urls=[]

siguiente='https://www.argenprop.com/casa-venta-localidad-capital-federal'
r=requests.get(siguiente)
casas = BeautifulSoup(r.content, 'html.parser')
dom= etree.HTML(str(casas))
can=dom.xpath('//li[@class="pagination__page"][3]/a')[0].text
can=int(can)
while True:
    r=requests.get(siguiente)
    if r.status_code==200:
        #Se extrae la información del html

        casas = BeautifulSoup(r.content, 'html.parser')
        dom= etree.HTML(str(casas))

        #Ubicación

        ubicacion = casas.find_all('p', attrs={"class":"card__address"})
        ubicacion=[i.text for i in ubicacion]
        #l_ubicacion.extend(ubicacion)

        #Descripción base

        descripcion_base = casas.find_all('h2', attrs={"class":"card__title"})
        descripcion_base=[i.text for i in descripcion_base]
        #l_descripcion.extend(descripcion_base)

        #Área

        area=dom.xpath('//ul[@class="card__main-features"]/li[1]/span')
```

```
        area=[i.text for i in area]
        #l_area.extend(area)

        #Dormitorios

        dormitorios=dom.xpath('//ul[@class="card__main-features"]/li[2]/span')
        dormitorios=[i.text for i in dormitorios]
        #l_dormitorios.extend(dormitorios)

        #Antigüedad

        antiguedad=dom.xpath('//ul[@class="card__main-features"]/li[3]/span')
        antiguedad=[i.text for i in antiguedad]
        #l_antiguedad.extend(antiguedad)

        #Precio

        precio = casas.find_all('p', attrs={'class':'card__price'})
        precio=[i.text for i in precio]
        #l_precio.extend(precio)

        #Url

        urls = casas.find_all('a', attrs={'class':'card'})
        urls = [i.get('href') for i in urls]
        urls=['https://www.argenprop.com'+i for i in urls]
        #l_urls.extend(urls)


        if (len(ubicacion)==len(area)) & (len(ubicacion)==len(dormitorios)) &␣
↪(len(ubicacion)==len(antiguedad)) & (len(ubicacion)==len(urls)):
            l_ubicacion.extend(ubicacion)
            l_descripcion.extend(descripcion_base)
            l_area.extend(area)
            l_dormitorios.extend(dormitorios)
            l_antiguedad.extend(antiguedad)
            l_precio.extend(precio)
            l_urls.extend(urls)


        ini=dom.xpath('//li[@class="pagination__page␣
↪pagination__page--current"]/span')[0].text
        ini=int(ini)
    else:
        break
    #print(ini,can)
    if ini==can:
```

```
        break
    siguiente='https://www.argenprop.com'+dom.xpath('//ul[@class="pagination␣
 ↪pagination--links"]/li[contains(@class,"-next")]/a')[0].get('href')
```

[43]:
```
print(len(l_descripcion),len(l_ubicacion),len(l_area),len(l_dormitorios),len(l_antiguedad),len
```

```
3620 3620 3620 3620 3620 3620 3620
```

[44]:
```
#Se crea el dataframe y se tratan las columnas correspondientes

df = pd.DataFrame({'Descripción':l_descripcion,'Ubicación':l_ubicacion, 'Área':
 ↪l_area, 'Dormitorios':l_dormitorios,'Antigüedad':l_antiguedad, 'Precio':
 ↪l_precio, 'Link':l_urls})
df.Precio=df.Precio.str.replace('\n','').str.replace(' ','').str.
 ↪replace('USD','USD  ').str.replace('.',',')
df.Ubicación=df.Ubicación.str.replace('\n','').str.strip(' ').str.strip('\n')
df.Área=df.Área.str.replace('\n','').str.strip(' ').str.strip('\n')
df.Antigüedad=df.Antigüedad.str.replace('\n','').str.strip(' ').str.strip('\n')
df.Dormitorios=df.Dormitorios.str.replace('\n','').str.strip(' ').str.
 ↪strip('\n')
df['Fecha_obs']=date.today()
```

[45]:
```
df
```

[45]:

	Descripción	Ubicación	Área	Dormitorios	Antigüedad
0	Casa 4 Ambientes Patio/ Galería/ Quincho sobre…	Pareja al 3800	200 m² cubie.	4 dorm.	18 años
1	Impecable casa de 4 ambientes en 3 plantas - G…	Argerich 3717	185 m² cubie.	3 dorm.	12 años
2	Casa - Villa Devoto	Mercedes al 4900	300 m² cubie.	4 dorm.	8 años
3	Venta de Casa 5 ambientes con oficina y deposi…	Benito Juárez al 2900	330 m² cubie.	4 dorm.	15 años
4	Venta - Casa 6 ambientes sobre lote propio de …	Gualeguaychu al 3014	164 m² cubie.	4 dorm.	20 años
…	…	…	…	…	…
3615	Casa - Parque Chacabuco	Víctor Martínez 1942/44	250 m² cubie.	3 dorm.	30 años
3616	Casa en VENTA - Caballito - Felipe Vallese 1000	Felipe Vallese al 1000	100 m² cubie.	3 dorm.	40 años
3617	CASAS - CASA - PALERMO VIEJO, CAPITAL FEDERAL	Pasaje Santa Rosa 2100	205 m² cubie.	3 dorm.	3 baños
3618	Venta Casa MULTIFAMILIAR, lote 8,66x20, CUB 35…				
3619	VENTA PH 3 AMBIENTES ALMAGRO CON TERRAZA PERMUTA				

```
3618            Galicia 1900  348  m² cubie.     5 dorm.    34 años
3619               Gallo 900  122  m² cubie.     2 dorm.    70 años

               Precio                                                Link  \
0        USD  690,000  https://www.argenprop.com/casa-en-venta-en-vil…
1        USD  370,000  https://www.argenprop.com/casa-en-venta-en-vil…
2        USD  595,000  https://www.argenprop.com/casa-en-venta-en-vil…
3        USD  430,000  https://www.argenprop.com/casa-en-venta-en-vil…
4        USD  350,000  https://www.argenprop.com/casa-en-venta-en-vil…
…              …                                                    …
3615         $145,000  https://www.argenprop.com/casa-en-venta-en-par…
3616  Consultarprecio  https://www.argenprop.com/casa-en-venta-en-cab…
3617     USD  495,000  https://www.argenprop.com/casa-en-venta-en-pal…
3618     USD  400,000  https://www.argenprop.com/casa-en-venta-en-vil…
3619     USD  249,999  https://www.argenprop.com/casa-en-venta-en-alm…

       Fecha_obs
0     2024-01-29
1     2024-01-29
2     2024-01-29
3     2024-01-29
4     2024-01-29
…          …
3615  2024-01-29
3616  2024-01-29
3617  2024-01-29
3618  2024-01-29
3619  2024-01-29

[3620 rows x 8 columns]
```

[ ]: