HOUSES TO SELL-EXERCICI WEBSCRAPING

Autors:

Roger Álvarez

Xavier Borrat

1. **Context.** Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació. Indicar l'adreça del lloc web.

Donada la creixen inestabilitat econòmica, increment de la inflació i la conseqüent pujada dels tipus d'interès, el sector immobiliari sembla estar a punt d'entrar en un canvi de cicle. Això genera dubtes i preocupacions a les persones que estan buscant comprar o vendre un habitatge.
És per això que sembla interessant poder obtenir informació sobre habitatges en venda per tal de poder seguir l'evolució d'aquests en el temps i per tan poder obtenir algunes respostes o ser capaços de poder preveure com evolucionar el sector.

Per fer-ho hem decidit fer web scrapping en un dels portals de venda de cases més important del país, Habitaclia.

Més en concret ens hem centrat en Barcelona ciutat per tal de poder simplificar el data set, amb la qual cosa l'enllaç de referència a sigut:

https://www.habitaclia.com/viviendas-barcelona.htm

El motiu d'aquesta selecció és que d'entre els portals estudiats d'on podríem obtenir la informació requerida, HABITACLIA ens ha semblat el que permetia fer el scrapping d'una manera més senzilla i ràpida.

2. **Títol.** Definir un títol que sigui descriptiu pel dataset.

El títol del data set és: Houses to Sell

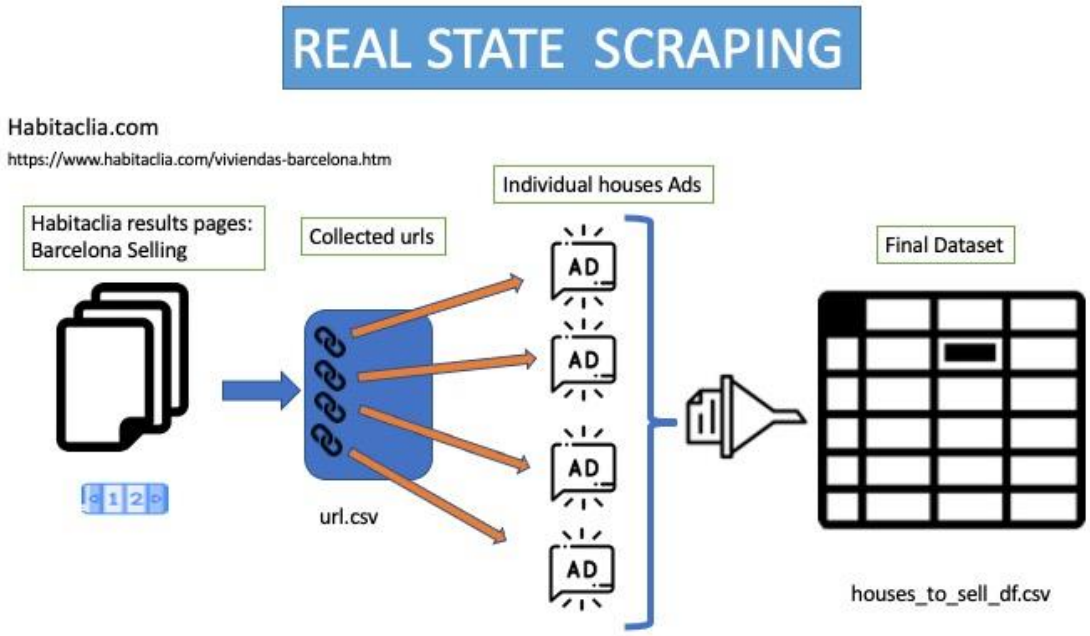
3. **Descripció del dataset.** Desenvolupar una descripció breu del conjunt de dades que s'ha extret. És necessari que aquesta descripció tingui sentit amb el títol escollit.

Les dades extretes i guardades en Houses to Sell són un conjunt de dades que ens permeten entendre alguns dels paràmetres més rellevants dels habitatges a la venda en la zona de Barcelona.

Paràmetres com preu de venda, superfície de l'habitatge o ubicació, entre d'altres, ens permetran poder fer un anàlisis sobre l'evolució de l'oferta d'habitatges en l'àrea de Barcelona en funció del temps.

Comentar també la intenció de que aquest data set sigui dinàmic. És a dir la idea és que de manera periòdica (cada dia, setmana, etc) es faci un run del main.py i la informació s'afegeixi al data set existent (especificant la data i hora). D'aquesta manera podrem estudiar com evolucionen els preus dels habitatge en funció del temps i podrem treure paràmetres com per exemple temps en que es tarda a vendre un habitatge en una zona determinada.

4. **Representació gràfica.** Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.



5. **Contingut.** Explicar els camps que inclou el dataset i el període de temps de les dades.

Product_id: ens indica el codi de l'habitatge. És la primary key de cada habitatge en el data set

Advert_id: és el codi de l'anunci en particular. La diferència entre aquest i l'anterior és que per exemple un habitatge pot estar anunciat per diferents agències, de manera que en aquest cas el product_id seria el mateix però el advert_id seria diferent.

Price: indica el preu de venda de l'habitatge.

City_name: és el nom de la ciutat on es troba l'habitatge. En el nostre cas sempre serà Barcelona ja que ens hem centrat en aquesta ciutat.

Zone: és el nom de la zona o el barri on es troba l'habitatge.

City_code: és el codi de la ciutat on es troba l'habitatge, en el nostre cas sempre serà 1930008. Aquest valor és redundant a City_name, però el considerem útil ja que en certes ocasions pot ser útil tenir el codi en format numèric.

Zone_code: és el codi de la zona o barri. És més útil que el Zone ja que aquest al ser una string pot tenir moltes diferències entre anuncis, mentre que el Zone_code és únic a cada zona el que el fa molt útil per filtrar o operar per zones.

Sqaure_meters: proporciona la superfície de l'habitatge en metres quadrats.

Bed_rooms: indica el número d'habitacions de l'habitatge.

Toilets: són el número de lavabos de l'habitatge.

Advert_url: és l'enllaç de l'anunci per si es vol consultar. Ens ha sigut molt útil per a poder comprovar que el data set era correcte.

Today's_date: en aquest paràmetre es fixa l'aspecte temporal. Ens indica la data i hora del moment de descarregar dels valors de l'habitatge.

Inicialment aquests són els paràmetres que guardarem en el data set, però destacant que es poden afegir i treure paràmetres en funció de les necessitats específiques.

6. **Propietari.** Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

Propietari

El propietari del conjunt de dades, obtingut utilitzant la llibreria whois de python, es habitaclia.com.

Projectes similars

Hem trobat projectes similars: https://github.com/davduran/spider/blob/master/spider.py

i https://github.com/Real-Estate-Scrapy els quals exploren les pàgines d'anuncis d'habitatges capturant informació molt útil per caracteritzar el mercat immobiliari i fins i tot construir eines predictives per estimar preus de vivenda i decidir si estan per sobre o per sota el preu de mercat.

Principis ètics i legals

Els autors d'aquest treball hem llegit els avisos legals publicats a la pàgina: https://www.habitaclia.com/hab_cliente/legal_aviso.asp

En aquest apartat d'avís legal s'especifica que es tracta d'un contingut de navegació pública. Així mateix especifica per als anunciants la informació ha de ser veraç. Considera propietat intel·lectual els títols, logos, figures, marques, software i base de dades del portal. Per tant prohibeix la utilització, modificació o explotació del contingut del portal amb excepció dels supòsits contemplats dintre els principis de bona fe. En aquest últim concepte és on pensem que estem legitimats per fer webscraping ja que la recerca acadèmica i ús docent sense interessos comercials entra dintre les bases legítimes incloses al principi de bona fe.

També hem descarregat i llegit el fitxer robots.txt per assegurar-nos que el nostre webscrapper no accedia a apartats no autoritzats marcats amb l'etiqueta disallow.

Fitxer: robots.txt

```
User-agent: *
Disallow: /hab_usuarios/registrocorreo.asp*
Disallow: /hab_usuarios/ajax/*
Disallow: /hab_inmuebles/ajax/*
Disallow: /dotnet/NotificacionesLiveListado/GetNotificacionesLiveListado*
Disallow: /dotnet/solicitud/vertelefono*
Disallow: /dotnet/solicitud/ValidarCaptcha*
Disallow: /dotnet/ficha/favrate*
Disallow: /dotnet/ficha/favcomment*
Disallow: /dotnet/ficha/translate*
Disallow: /*.txt$
Allow: /robots.txt
Allow: /app-ads.txt
Allow: /ads.txt
Disallow: /*habsrc=
Disallow: /*list=
Disallow: /*contactar.htm
Disallow: /q/
Disallow: /*/q/
Disallow: /*listainmuebles.htm
Disallow: /*ady=
Disallow: /*z=
Disallow: /*fotomode=
Disallow: /*codProv=
Disallow: /*codPob=
Disallow: /*openmenu=
Disallow: /*subtipinm=
Disallow: /*coddists=
Disallow: /*compartirApp=
```

nota: contingut del fitxer robots.txt de pàgina habitaclia.com

Pensem que la nostra pràctica s'ajusta als termes i condicions i es respecta la propietat intel·lectual on hem verificat les condicions d'ús, només hem rastrejat informació pública sense causar danys i sense cap interés comercial.

7. **Inspiració.** Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

L'interès per aquest conjunt de dades és molt evident. Contenen informació molt extensa del mercat immobiliari de Barcelona ciutat amb moltes característiques de cada habitatge que en el seu conjunt ens permet modelar i fer-nos una idea de les relacions entre característiques i el cost.

Pot ser útil tant per al comprador com per al venedor d'immobles ja que després de modelar les dades es pot introduir qualsevol habitatge per fer una estimació de preu i així veure si aquest habitatge està per sobre o per sota del preu de mercat. O de manera més simple es pot buscar la mitja de preu per m^2 de cada zona o el nombre d'habitatges al mercat per zona. Altres aplicacions podrien ser buscar la característica que més pes tenen per definir el preu i també si es repeteix l'scraping en diferents instàncies del temps, es pot detectar el temps que es tarda a vendre l'immoble i així ajustar encara més el model amb la informació de temps fins a venta.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció. Exemples de llicències que poden considerar-se:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Altres (especificar quina).

Apache License 2.0: Llicencia permisiva les condicions de la qual exigeixen la conservació dels drets d'autor i avisos de llicència. Els contribuients proporcionen una concessió expressa dels drets de patent. Les obres amb llicencia, les modificacions i les obres de major envergadura poden distribuir-se sota les diferents condicions i sense codi.

9. **Codi.** Codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

- El codi haurà de situar-se a la carpeta ***/source*** del repositori.

Link al repositori: https://github.com/xborrat/home_scraping

Hem creat tres fitxers a la carpeta source:

main.py: conté els paràmetres a modificar per variar el nombre de pàgines a explorar i la crida a les funcions contingudes en altres fitxers(advert_scraper.py i url_collector.py)

url_collector.py : Conté la funció page_link_collector que cerca les diferents url contigudes en les pàgines de resultats després de consultar habitatges en venda de barcelona. Genera un document amb tots els url capturats.

advert_scraper.py : Llegeix el document de url capturades i n'extreu tota la informació necessària per caracteritzar una vivenda. També escriu la informació de cada vivenda en una fila diferent d'un document csv (house_to_sell_df.csv)

● S'han d'indicar les llibreries i versions utilitzades. P. ex., en Python

```
pip3 freeze > requirements.txt
```

Fitxer: requirements.txt

```
aiohttp==3.8.1
aiosignal==1.2.0
async-timeout==4.0.2
attrs==22.1.0
beautifulsoup4==4.11.1
certifi==2022.9.24
charset-normalizer==2.1.1
frozenlist==1.3.1
future==0.18.2
idna==3.4
multidict==6.0.2
python-whois==0.8.0
requests==2.28.1
soupsieve==2.3.2.post1
typing_extensions==4.4.0
urllib3==1.26.12
yarl==1.8.1
github==1.2.7
```

● Al document PDF s'han de comentar els aspectes més rellevants sobre com el codi realitza el procés de recol·lecció de dades, quines dificultats presenta el lloc web triat, i com les heu resolt.

L'esquema global del programari és un script on es pot determinar l'extensió de la búsqueda en forma de nombre de pàgines que contenen els resultats d'anuncis d'habitatges en venta a barcelona. Aquest mateix script(main.py) crida a dues funcions de forma correlativa. La primera recull les

direccions URL a cada anunci individual de cada pàgina(url_collector.py) i la segona explora cada una d'aquestes URL per extreuren l'informació rellevant per la pràctica(advert_scraper). Aquesta última informació es guarda en un document csv.

Per no ser bloquejats i a més espaiar les peticions HTTP per no saturar la pàgina hem introduit al codi una modificació de l'user agent a la capcelera de la intrucció get depenent de la classe requests i hem afegit una ordre delay respectivament.

La principal dificultat trobada fou que uns pocs anuncis del total (un de cada 100) tenien una estructura diferent a la resta i a més aquesta variació tampoc era consistent per això ho hem gestionat aixecant una excepció i per tant no registrem aquests anuncis anòmals.

10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset (https://doi.org/...). El dataset també haurà d'incloure's a la carpeta **/dataset** del repositori.

DOI: 10.5281/zenodo.7338399

11. **Vídeo**.

Link al video:

https://drive.google.com/file/d/1Oc549dD_KJAgwC4HooybSW_RDbKZpwXn/view?usp=share_link

Contribució	Signatura
Investigació prèvia	XBF, RAC
Redacció de respostes	XBF, RAC
Desenvolupament del codi	XBF, RAC
Participació al vídeo	XBF, RAC