Pràctica 1 – Tipologia i cicle de vida de les dades

Assignatura: M2.951 Data: Novembre 2023

Autors	Àlex Tort – atort9@uoc.edu Laura Guerra – lguerrari@uoc.edu
Lloc web escollit	https://pisos.ad/
Repositori	https://github.com/Laura-Guerra/WebScraping_TCVD
Enllaç DOI Zenodo	https://doi.org/10.5281/zenodo.10112197
Video de presentació	https://drive.google.com/file/d/1kdzwZ23TQ6NLMVpBm1y-rIJN3sL_zJfT/view?usp=sharing

Resolució dels apartats

1. CONTEXT

Durant els últims anys el mercat immobiliari andorrà ha observat un augment notable en la demanda d'habitatges, tant de compra com de lloguer, degut al creixent interès de residents estrangers pel país. A part dels avantatges fiscals, altres incentius com la baixa taxa d'atur, l'entorn o la qualitat de vida han convertit a Andorra en un dels països amb un nombre més elevat de nous residents. Aquesta increment de residents ha fet augmentar la demanda d'habitatges, fins al punt en què aquesta supera l'oferta disponible.

Degut a la geografia del país, envoltat per muntanyes, el desenvolupament immobiliari està restringit físicament. Aquesta limitació territorial i l'augment en la demanda esmentada ha donat lloc a un mercat molt competitiu amb preus constantment a l'alça; convertint l'accés a l'habitatge en un problema de dimensions cada vegada majors. Moltes persones, tant de nacionalitat andorrana com potencials nous residents, s'han vist obligats a buscar alternatives que sovint impliquen buscar un habitatge a localitats properes però fora del país, com per exemple La Seu d'Urgell.

Per aquests motius hem considerat que disposar de dades detallades del mercat immobiliari andorrà, tant de venta com de lloguer, pot resultar molt útil. En aquest aspecte, obtenir dades d'una pàgina web immobiliària permet obtenir dades per zona geogràfica, tipus d'habitatge i altres característiques específiques a partir de les quals es pot estudiar la distribució de preus, determinant, per exemple, els factors que influeixen en la diferència d'aquest. També es pot estudiar quines parròquies estan experimentant un major creixement, i determinar si cal dur a terme alguna acció en conseqüència.

Així, la plataforma escollida ha estat www.pisos.ad , un dels portals immobiliaris andorrans amb una major oferta de propietats. Tal com anuncien al seu portal, disposen de més de 3.000 propietats, publicades bé per particulars bé per una de les 34 immobiliàries que utilitzen la plataforma. A part, la plataforma permet cercar immobles per parròquia, facilitant així la classificació d'aquests en funció de la seva ubicació. L'elevada oferta, així com la popularitat de la pàgina entre els residents del país, ens ha conduit a considerar aquest portal com el més

indicat per dur a terme l'estudi proposat en aquesta introducció i que es detalla en més profunditat als apartats posteriors.

2. TÍTOL

El títol escollit pel nostre data set és "PREUS PROPIETATS ANDORRA"

3. DESCRIPCIÓ DEL DATASET

Per tal de dur a terme l'estudi proposat a l'apartat anterior, s'ha recopilat informació de totes les propietats publicades al portal immobiliari a data del 9 de novembre de 2023, classificant-les en funció de la tipologia de transacció (venda o lloguer) i per parròquia. Si bé no es pot garantir que les dades actuals són estrictament d'habitatges, durant el procés de neteja del conjunt de dades es pot garantir el filtratge d'aquelles que no es corresponen amb aquesta categoria a través de la variable referent al nombre d'habitacions.

4. REPRESENTACIÓ GRÀFICA

5. CONTINGUT

Cada registre del conjunt de dades correspon a un anunci publicat al portal. No obstant, no necessàriament ha de tractar-se d'una propietat en particular, ja que diferents agències poden publicar anuncis de la mateixa propietat. Així doncs, per cada registre, es recullen les dades següents:

- **Price**: preu establert a l'anunci. En cas de tractar-se d'una propietat de lloguer, el preu correspon al preu mensual d'aquesta. El format actual de la variable és el valor en format text, amb punts als milers i en €.

- **Area**: superfície de la propietat, em m2. De nou, actualment es tracta d'una variable textual composta pel valor de la superfície de la propietat i les unitats corresponents.

- **Bedrooms**: nombre d'habitacions de la propietat. En cas de no tractar-se d'un immoble, aquesta variable té com a valor "0 Habitacions".

- **Parking**: variable categòrica binària que indica si la propietat disposa de garatge. Els seus valors són "Inclòs" i "No Inclòs".

- **Features**: llista de característiques de la propietat.

- **Agency**: nom de l'agència immobiliària que ha publicat l'anunci.

- **Id**: identificador numèric únic de l'anunci, que es correspon amb la referència del portal.

- **Type**: tipologia de la transacció anunciada. Els possibles valors de la variable són "venda" i "lloguer".

- **Zone**: parròquia on es troba ubicada la propietat. Els possibles valors són "ordino", "canillo", "encamp", "andorra-la-vella", "sant-julia-de-loria", "escaldes-engordany" i "la-massana".

- **URL**: enllaç de l'anunci.

- **Timestamp**: data i hora de l'extracció. Degut a l'elevada demanda, és possible que part dels anuncis (sobretot els de lloguer) deixin d'estar publicats al portal al cap d'uns dies i per tant no es pugui accedir de nou a la informació. Per aquest motiu s'ha considerat rellevant indicar el període al qual pertanyen les dades.

El conjunt de dades pot trobar-se a https://doi.org/10.5281/zenodo.10112197.

6. PROPIETARI

La pàgina web exposa el següent text:

"Avís Legal s'informa que www.pisos.ad és un domini de l'empresa ANTONI CAPELLA FERNANDEZ (en endavant "PISOS.AD") amb domicili social al Av de les Escoles, 31 – 5º5ª -AD700- ESCALDES-ENGORDANY (PRINCIPAT D'ANDORRA), i correu electrònic: info@pisos.ad

Inscrita al Registre de Comerç del Principat d'Andorra amb número 920418-Z i amb número de Registre Tributari F-033235-T"

Després de verificar els termes i condicions, així com el fitxer robots.txt de PISOS.AD, s'ha pogut determinar que la recopilació de dades específiques com llistats d'immobles, preus, ubicacions i característiques generals està permès sempre que no s'infringeixin els drets de propietat intel·lectual de la plataforma, no es recopilin dades personals dels usuaris ni s'impedeixi el funcionament normal del lloc web.

Degut a l'absència d'anàlisis anteriors específics sobre el web scraping de dades de PISOS.AD, s'ha realitzat una recerca de casos similars en altres plataformes immobiliàries per identificar quines són les dades d'interès i garantir que el nostre procés s'ajusta a les normatives legals vigents. Exemples d'anàlisis en webs similars:

- Evolució del preu de la l'habitatge a Andorra: Indomio

- Preu de la habitatge a Andorra per m2: RealAdvisor

- Evolució del preu de la habitatge a Barcelona: Idealista

7. INSPIRACIÓ

Tal i com s'ha introduït en el primer apartat d'aquesta memòria, l'objectiu de l'anàlisi de les dades extretes del portal pisos.ad és obtenir una visió més detallada del mercat immobiliari andorrà.

Un dels primers objectius de l'anàlisi de les dades obtingudes és determinar la situació actual de la oferta i la demanda, i per tant, de la disponibilitat d'immobles. A part d'un primer estudi

generalitzat, també es pot determinar la distribució d'oferta per parròquies, per tal d'observar si la situació és similar a totes les parròquies o hi ha més oferta i/o demanda en funció de la localització dels immobles. L'estudi de la distribució de la oferta també es pot realitzar en funció de la tipologia de la transacció, és a dir, si la propietat és de lloguer o està en venta. Degut al context actual de pujades dels tipus d'interès, i conseqüentment les comissions de les hipoteques, seria interessant contrastar si s'ha observat una disminució en la demanda d'habitatges de compra i en conseqüència un augment en la demanda de lloguers. Finalment, seguint amb aquesta línia d'estudi, en un futur es podria repetir l'extracció i compara com ha evolucionat l'oferta immobiliària, tant per observar-ne la tendència com per determinar si aquesta es veu subjecta a estacionalitat (en el cas dels lloguers), fet molt probable degut a l'increment d'ocupació laboral i del turisme en temporada.

Una altra dimensió rellevant que volem explorar és la distribució de preus en funció de la ubicació de l'immoble, així com de les característiques d'aquest (nombre d'habitacions, superfície, garatge...), determinant quin pes pot arribar a tenir cada característica a l'hora d'establir els preus. En cas de poder assolir aquest objectiu, es podria determinar quins immobles estan per sobre del preu que els correspondria, i quins per sota. En cas que aquest conjunt de dades i el seu estudi estigués disponible al públic general, els inversors el podrien utilitzar per cercar oportunitats de negoci, si bé aquest fet agreujaria la situació descrita a la introducció de la memòria. Per contra, aquestes dades també podrien utilitzar-se per determinar quines zones estan més tensionades i intentar buscar solucions a través de la implementació de noves legislacions addicionals a la legislació actual de congelació dels preus de lloguer o de taxació de les inversions estrangeres.

Finalment, es pretén estudiar el nombre d'immobles publicats per cada immobiliària amb la intenció d'establir si hi ha lliure competència al sector o pel contrari aquest es troba monopolitzat. En cas que el conjunt de dades permetés determinar dos registres duplicats a través de les característiques del conjunt, però publicats per immobiliàries diferents, també es podria determinar si els preus per un mateix immoble publicat per dues immobiliàries diferents és el mateix, o si per contra, una d'elles té un preu més elevat.

Si bé s'han realitzat estudis similars, com en l'anàlisi de RealAdvisor, considerem que l'estudi proposat aporta informació addicional, com podria ser el fet de que la propietat disposi de pàrquing, algunes de les característiques que es recullen al portal immobiliari, o tot l'estudi relacionat amb les immobiliàries que gestionen els anuncis de les propietats.

8. LLICÈNCIA

- S'ha de reconèixer l'autoria de manera apropiada, proporcionar un enllaç a la llicència i indicant si s'han realitzat canvis al material original.

- En el cas de modificar o crear a partir del material original, s'han de distribuir sota la mateixa llicència que l'original.

Per més informació sobre la llicència visitar Creative Commons.

9. CODI

El repositori amb el codi utilitzat en aquest projecte és el següent:

https://github.com/Laura-Guerra/WebScraping_TCVD

El codi està estructurat en els fitxers següents:

- constants.py: fitxer on es defineixen les diferents constants necessàries al llarg del procés (enllaços base, classes, filtres, rutes...).
- utils_files.py: fitxer on es defineixen les funcions necessàries per generar i llegir els fitxers csv.
- utils_url.py: fitxer on es defineix la funció que genera els diferents enllaços sobre els quals s'itera.
- utils_scraping.py: fitxer on es defineixen les funcions d'obtenció dels enllaços i del contingut d'aquests.
- main.py: codi principal a través del qual es criden les diferents funcions necessàries per realitzar el procés de scraping.

En executar el codi principal des del fitxer main.py, s'executa la funció main(). En primer lloc, aquesta mostra per pantalla el User-Agent utilitzat tant per Selenium com l'establert per BeautifulSoup (BS4). En el cas de BeautifulSoup l'és "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/119.0.0.0 Safari/537.36" i en el cas de Selenium depèn del navegador de l'ordinador que executa el programa. Seguidament es crida la funció get_zone_urls(), definida en un dels fitxers auxiliars, que obté els diferents enllaços corresponents a les diferents parròquies i tipologies de transacció (venda o lloguer).

La principal dificultat que hem trobat en aquest procés d'extracció de dades ha estat obtenir els enllaços de filtre a les diferents parròquies. Aquest filtre es realitza mitjançant el widget d'un mapa que es carrega dinàmicament. El problema amb els elements que es carreguen d'aquesta manera és que no es poden detectar amb les llibreries requests i BeautifulSoup, ja que només accedeixen al contingut present en l'HTML a l'inici de la càrrega de la web.

Per a solucionar-ho hem utilitzat la llibreria Selenium, que a diferència de requests i BeautifulSoup, simula una navegació real en un navegador, el que permet que els elements dinàmics tinguin temps suficient per carregar-se completament. Relacionat amb això mateix, un altre problema que ens vam trobar va ser que, inicialment, utilitzàvem la funció "visibility_of_all_elements_located" de Selenium, amb la qual tampoc trobàvem el llistat complet de URLs per zona. Això es deu a que aquesta funció només pot detectar elements que ja estan carregats en el DOM i que són visibles a la pàgina web. Per solucionar-ho, vam substituir aquesta funció per presence_of_all_elements_located de Selenium, que pot detectar tots els elements carregats en el DOM, independentment de si són visibles a la pàgina o no.

A continuació s'executa la funció get_house_urls(zone_urls), que accedeix als enllaços de les diferents parròquies, tant per a propietats en venda com en lloguer. Iterant a través de la paginació de cada url, el programa desa en un fitxer CSV les URL de les redireccions a les pàgines de detall de cada propietat, juntament amb un ID associat que s'obté de la pròpia url, la parròquia on es troba la propietat i si està en venda o lloguer. Aquesta iteració inclou una pausa entre sol·licituds d'entre 1 i 3 segons per ajustar-se a les bones pràctiques del web scraping.

Finalment, es llegeix els diferents enllaços i informació continguda al CSV generat i s'executa la funció generate_df, que retorna el conjunt de dades definitiu. Aquesta funció processa de forma concurrent els elements del conjunt de dades del CSV generat amb anterioritat i per cada registre es reconstrueix la URL completa a través de la funció concatenate_url i n'obté el contingut a partir de get_page_content. Un cop obtingut el contingut de l'enllaç, es seleccionen les característiques d'interès (preu, superfície, nombre d'habitacions, aparcament, llista de característiques i informació de l'agència) i es desen en un diccionari. Els diccionaris obtinguts per cada iteració s'afegeixen a una llista que posteriorment es transforma en un dataframe i es desa en un fitxer CSV. L'execució d'aquest últim mòdul es realitza de forma concurrent ja que es va comprovar que, tot i realitzar les sol·licituds sense deixar una pausa entre elles, no es sobrecarregava el servidor. Així doncs, s'ha optat per optimitzar el procés i reduir el temps d'execució.

10. DATASET

El data set obtingut en aquest projecte és el següent:

https://doi.org/10.5281/zenodo.10112197

11. VIDEO

El vídeo de presentació del nostre projecte es pot visualitzar al següent enllaç:

https://drive.google.com/file/d/1kdzwZ23TQ6NLMVpBm1y-rIJN3sL_zJfT/view?usp=sharing

Contribucions	**Signatura**
Investigació prèvia	A.T. , L.G.
Redacció de les respostes	A.T. , L.G.
Desenvolupament del codi	A.T. , L.G.
Participació al vídeo	A.T. , L.G.