Victor Baena Chamizo
Arnau Muns Orenga

Tipologia i cicle de vida de les dades

Pràctica 1 - Web Scraping

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

No és cap secret que en la societat actual els fills han de marxar eventualment de casa els pares. Aquest procés d'independència familiar sovint es complica a causa de diferents motius. Per exemple, els elevats preus dels lloguers o les compres, les feines precàries a què normalment hem d'accedir els joves (que generalment no permeten pagar una entrada d'una casa molt elevada i obliguen al jovent a haver de compartir pis si es lloga), etc. Malgrat això el procés d'independència és un fet que esdevé natural en totes les famílies i pel qual els integrants d'aquest grup estem passant, motiu pel qual ens semblava adient aprofitar aquest context per relacionar-lo amb la pràctica actual.

En aquest context s'ha decidit, per aquesta pràctica, recollir informació del principal portal immobiliari espanyol en línia (idealista.com) sobre totes les cases i pisos en venda d'una població, hem triat la ciutat d'Igualada per motius pràctics ja que un membre estaba considerant l'adquisició d'un pis i així ens permatia també donar-li més utilitat, amb l'objectiu de poder crear una base de dades unificada que permeti una anàlisi en profunditat sobre els factors que determinen el preu de compra d'un pis, així com per intentar identificar quins pisos o cases serien més atractius per comprar en relació amb la situació de la zona en la qual es troben. En definitiva es tracta d'un procés de recollida de dades que pot ajudar als integrants d'aquest equip a fer una cerca més exhaustiva i analitica de l'oferta actual de pisos i cases i, amb una mica de sort, ajudar-nos a descobrir la millor opció de compra.

És evident, doncs, per què el lloc web escollit proporciona informació sobre compra/venda/lloguer de pisos i cases. Idealista.com, tal com es defineix la mateixa empresa, és el portal immobiliari online espanyol més gran del sector. És una companyia fundada l'any 2000 que ofereix a través d'internet els serveis de portal immobiliari a Espanya, Italia i Portugal. Tanmateix, aquest servei no és l'únic que ofererreixen perquè també tenen una línia de negoci enfocada a la negociació d'hipoteques, així com serveis de dades (per exemple, basant-se en un pis concret t'ofereixen un informe analític sobre la situació d'aquell pis amb preus entre 11.99€ i 20€ depenent de la zona).

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

Aquest conjunt de dades porta per títol “característiques_pisos_igualada.csv”.

L’elecció d'aquest títol i aquest enfoc, és que ens trobem en un punt on la cerca de pis pot ser un problema i hem pensat que podem contribuir a donar-hi una solució o com a mínim que tingui una certa utilitat més enllà de la realització d’aquesta.

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.**

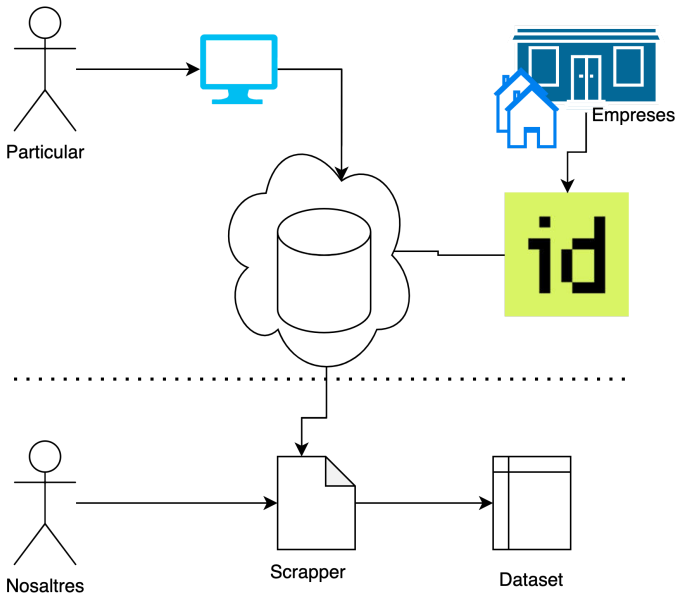
Tal i com dóna a entendre el títol, aquest conjunt de dades està format per les diferents característiques dels pisos d’una zona determinada. Aquestes característiques inclouen, entre d’altres, el preu o els m2 del pis o casa concret. Aquestes característiques es detallen en les següents preguntes.

En aquest conjunt de dades cada observació fa referència a un pis o casa diferent, presentant-se per columnes totes les característiques prèviament esmentades. Quant al factor temporal cal destacar que aquest conjunt de dades és de tall transversal, això és, captura les característiques dels pisos i/o les cases en un moment determinat del temps (Abril 2022) i no presenta cap representació temporal dels mateixos. Les unitats de mesura de les característiques varien en funció de cada variable i es detallaran en l’apartat corresponent.

En aquest primer estadi del projecte les dades no han passat per cap procés de preprocessament o neteja. És per això que les dades encara poden presentar inconsistències i el format no és necessàriament el millor per un anàlisi directe. El preprocessament d’aquest conjunt de dades es planteja en el segon estadi d’aquest projecte. El format del conjunt de dades és un fitxer .CSV per tal de facilitar la seva visualització i posterior tractament i anàlisi.

4. **Representació gràfica.**

Donat que el nostre projecte consisteix a l'extracció de dades d'una pàgina de compravenda de pisos (Idealista) el que hem pensat es en establir un fluxe com el que mostra la imatge a continuació.



Donat un particular que busca vivenda i I les empreses que es dediquen a promocionar-les, Idealista en aquest cas els posa en contacte mitjançant la seva plataforma web, on desar al núvol totes les vivendes.

A partir d'aquest punt es on entrem nosaltres, ja que mitjançant el nostre scraper, treiem els pisos i generem el nostre dataset un cop netejades les dades.

5. **Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.**

Per la realització d'aquesta pràctica hem extret un seguit de dades de la web d'habitatges. Una tasca que hem realitzat ha sigut filtrar la informació que hem cregut útil pel nostre dataset, això ens ha portat a recollir les següents variables per un total de 307 observacions o pisos:

Nom variable	Descripció	Unitat de mesura	Tipus variable
id	Identificador del pis	-	Numèrica
construidos_m	Metres quadrats construïts	m2	Numèrica
utiles_m	Metres quadrats útils	m2	Numèrica
habitaciones	Número d'habitacions	-	Numèrica
baños	Número de lavabos	-	Numèrica
terraza	Número de terrasses	-	Numèrica
balcon	Existència de balcó	-	Boolean (SI/NO/NA)
estado	Estat del pis o casa	-	Categòrica politòmica
orientacion	Orientació respecte els punts cardinals	-	Categòrica politòmica
garaje	Existència de garatge	-	Boolean (SI/NO/NA)
año_construido	Any de construcció de l'edifici on està el pis o la casa	Any	Numèrica
calefaccion	Existència de calefacció de qualsevol tipus	-	Boolean (SI/NO/NA)
planta	Número de planta on es troba el pis	-	Numèrica
ascensor	Existència d'un ascensor a la finca	-	Boolean (SI/NO/NA)
aire_acondicionado	Existència d'aire condicionat al pis	-	Boolean (SI/NO/NA)
zonas_verdes_o_jardin	Existència de zones verdes o jardí a la finca o al pis/casa	-	Categòrica politòmica

ciudad	Ciutat on es troba el pis o la casa	-	Text
zona	Zona dins la ciutat on es troba el pis o la casa	-	Text
etiqueta_energetica	Lletra de la certificació energètica del pis o la casa	-	Categòrica politòmica
precio	Preu de venda	€	Numèrica
precio_anterior	Preu de venda anterior a la rebaixa (si n'hi ha)	€	Numèrica
titulo_anuncio	Títol del anunci a idealista.com	-	Text

El període de temps de les dades es força actual, ja que la propia web s'encarrega d'anar actualitzant els anuncis i els professionals que oferten pisos i cases també les acostumen a actualitzar sovint (per sortir més amunt a la pàgina). La data màxima d'extracció de les dades va ser el dia 8 d'Abril de 2022, amb la qual cosa el nostre conjunt de dades presenta informació dels pisos i les cases en venda a Igualada en aquell moment del temps. No s'ha considerat extreure informació del mateix pis o casa al llarg del temps ja que considerem que és més interessant elaborar un anàlisi de tipus transversal en un moment del temps que fer un anàlisi de seguiment de diferents pisos al llarg d'un període.

La recol·lecció de les dades s'ha dut a terme mitjançant un scrapper que s'ha construït en python amb les llibreries *requests* i *BeautifulSoup*. La idea del procés d'extracció de les dades es fonamenta en un procés iteratiu.

Un dels principals reptes que vam tenir és el de la paginació. La informació dels pisos d'idealista.com està distribuïda a través de diferents pàgines dins la seva web. Cada pàgina conté un seguit d'anuncis de pisos que condueixen a la informació del pis concret. Per aquest motiu, en primer lloc es captura la llista de totes les pàgines de pisos per Igualada (una localització determinada) amb l'objectiu de conèixer quin és el número màxim de pàgines sobre les quals hem de cercar la informació dels pisos.

Una vegada capturat aquest nombre de pàgines mitjançant una funció que hem creat, s'utilitza un procés iteratiu sobre aquest llistat de pàgines. Per cada una de les pàgines s'aplica un altre funció que treu totes les URL dels pisos anunciats en la pàgina que s'està mirant actualment. Una vegada capturada la llista de pisos per una pàgina "i" es procedeix a fer un altre procés iteratiu per cada un dels pisos. Per cada un dels pisos de la pàgina "i" s'aplica el procés de parsejat de la informació. Es capturen totes les dades que s'han considerat rellevants sobre

aquell pis i s'afegeixen a una llista on es van acumulant totes les dades dels diferents pisos. El resultat d'aquest doble procés iteratiu és una llista parsejada de la informació de cada pis que posteriorment es transforma a format .csv per l'elaboració d'aquesta pràctica.

Val a dir que un altre repte que ha calgut superar és el d'evitar el bloqueig de les nostres URL per part d'idealista.com. Per aquest motiu es va decidir separar cada una de les peticions del nostre scraper entre uns 50 i 80 segons de manera aleatòria per tal que idealista no bloquegés les nostres IPs al detectar-nos com a bots.

6. **Agraïments. 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.**

El propietari del lloc web que s'ha scrapejat en aquesta pràctica és una empresa que té com a nom legal Idealista, S.A.U. Aquesta empresa gestiona la web idealista.com i totes les apps relacionades. La seva seu social està a Madrid i és una empresa registrada en el registre mercantil de Madrid amb NIF A82505660. Aquesta empresa pot ser contactada a través del seu lloc web en el següent enllaç per qualsevol pregunta o comunicació que es vulgui establir amb ells: https://www.idealista.com/info/contacto.

Tal i com indica idealista en els seus Termes i Condicions Generals d'Idealista (1), especificament en l'apartat "¿Qué puedes y no puedes hacer en idealista?", no està permès fer servir mecanismes automàtics per copiar o extreure el seu contingut. Més endavant en el mateix document es detalla que no està permès "accedir, controlar o copiar qualsevol informació inclosa en la web fent servir qualsevol tipus de robot, spider, scraper o un altre mitjà automàtic o procés manual per qualsevol propòsit sense el nostre permís exprés i per escrit". És per això que, previ a realitzar scraping i amb l'objectiu d'actuar d'acord amb el principis ètics i legals en el context d'aquest projecte, s'ha decidir contactar directament amb idealista.com per tal de demanar permís per poder accedir a la informació de la seva pàgina web mitjançant un scraper per aquesta pràctica. Idealista va accedir sense cap problema a permetre'ns realitzar scraping sobre la seva web al tractar-se d'un projecte acadèmic i sense ànim de lucre tot i que van presionar bastant perquè demanéssim accés a la seva API pública. Aquest accés no es va demanar perquè està limitat, segons ells expliquen, a 500 peticions cada dos mesos i s'ha considerat que són massa poques per la realització d'aquesta pràctica.

Això no sempre s'aplica ja que d'altres empreses com Habitaclia van rebutjar colaborar. Tot i que es va contactar via mail, van argumentar que el seu portal era destinat a la consulta particular per la cerca de pisos i no pas per l'extracció d'informació, i que els permetre aquestes accions els suposaria una càrrega al seus servidors.

Francament creiem que és senzillament una excusa per no permetre'ns extreure informació ja que imagino que els serà de gran valor i creuen que volem treure profit (apart del acadèmic). Això ens ha portat a la conclusió que les empreses grans ja son conscients del valor de les

seves dades i no totes son tan obertes a proporcionar-les encara que sigui per a activitats acadèmiques, tot i que com hem comentat Idealista si ens va oferir formes de col·laboració.

Quant a altres anàlisis similars sobre el mateix lloc web d'idealista, aquests es poden trobar fàcilment a internet:

A. https://www.youtube.com/watch?v=2UyJv5oe570&t=1352s En aquesta sèrie de videos, l'autor Miguel Ángel Gisbert intenta crear un scraper molt semblant al que s'ha creat per aquesta pràctica però amb lleugeres diferències. En aquest cas ell està interessat en extreure la informació dels pisos a partir d'una cerca personalitzada basada en el codi postal. També presenta la seva solució al problema de la paginació però considerem que és més ineficient que la nostre ja que requereix una cerca exhaustiva per totes les pàgines per extreure la llista total de pàgines. Tampoc planteja cap solució al problema del bloqueig d'IP per la qual cosa acaba tallant el vídeo perquè li bloquegen la IP en directe…
B. https://www.youtube.com/watch?v=4Tv73KuqgVo&t=4692s&pp=ugMICgJlcxABGAE%3D En aquest cas l'autor del projecte utilitza la llibreria de Python *Scrapy* per crear un codi que extreu informació dels pisos d'idealista també a partir d'una cerca personalitzada en base al codi postal. L'autor comenta que la seva empresa ha acabat abandonant el projecte a causa de les fortes mesures anti-scraper que té Idealista.
C. https://github.com/David-Carrasco/Scrapy-Idealista En aquest altre projecte l'autor, David Carrasco, presenta un codi Python ja preparat per a que altres usuaris puguin extreure dades d'idealista.com d'una manera senzilla en base a una URL. Aquest projecte presenta un punt interessant i és que l'autor l'ha creat pensant que es pugui executar en un entorn de màquines virtuals com Docker.

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.**

Com ja s'ha anat comentant al llarg de la pràctica, considerem que aquest conjunt de dades és molt interessant per gent jove com nosaltres que estigui buscant pis ja que aporta informació de totes les cases i pisos en venta en un moment concret del temps en una ciutat determinada, en aquest cas Igualada. Per inspirar-nos hem intentat imaginar quines informacions serien útils a l'hora de buscar el nostre primer pis. Alguns exemples de preguntes a les quals es pretén donar resposta amb aquest conjunt de dades són les següents:

<u>Preguntes relacionades amb els pisos en si</u>

- Quins són els 5 pisos més barats/assequibles de tota Igualada?
- Quins són els pisos més barats en €/m2 de tota Igualada?
- Quin és el pis més car i quin el més barat?
- Quin és el pis que està millor equipat?
- Quin és el pis amb més habitacions? amb més lavabos? etc
- Quin és el pis amb una major ràtio d'equipament/€?

Preguntes relacionades amb la zona

- Quina zona té més anuncis de pisos en venda?
- Quina és la mitjana de preu dels pisos segons les diferents zones de la ciutat?
- Quina és la zona més barata i quina la més cara (en mitjana)?
- Quina és la zona amb més dispersió de preus?

Preguntes generals

- Quin seria el millor pis per comprar en base a les seves característiques si es disposa d'un pressupost limitat de X €?
- Quina seria la recomanació de compra per diferents pressupostos?
- Quina seria la recomanació de compra per un pressupost infinit segons la zona?

En general la idea de projecte d'aquest conjunt de dades és la de poder construir un model predictiu del preu de venda d'un pis en base a les seves característiques. Aquest model, evidentment, vindrà precedit per un anàlisi descriptiu que pot incloure preguntes com les que s'han exemplificat en el paràgraf anterior. Val a dir que en els anàlisis similars que s'han presentat en l'apartat anterior l'objectiu era pràcticament el mateix. Tot i que els projectes esmentats es centren única i exclusivament en el procés d'scraping i extracció de la informació, els seus creadors mencionen que l'objectiu és el de crear un model predictiu del preu de venda en base a les característiques dels pisos per tal de poder predir, per un pis nou que es posa a la venda, quin hauria de ser el seu preu en base a altres pisos semblants que ja estiguin ofertats per internet.

8. **Llicència.**

Per la realització d'aquesta pràctica ens hem decantat per la següent llicència:

Released Under CC BY-SA 4.0 License (ús comercial)

Realment hem triat de forma hipotética, però creiem que per un ús acadèmic i si fos un treball seriós amb opcions comercials seria la més adient:

Aquesta llicència permet a altres barrejar, retocar i basar-se en el nostre treball fins i tot amb finalitats comercials, sempre que ens nomen en els credit i publiquin les seves noves creacions sota els mateixos termes.

9. **Codi. Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.**

https://github.com/UOC-Victor-Baena/Practica1

10. **Dataset. Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.**

DOI: 10.5281/zenodo.6448264
Enllaç: enllaç ZENODO

11. Vídeo

Drive

https://drive.google.com/file/d/1JWcvr2d7UrkyvKTSZ1XRoYqVIo-S_dk2/view?usp=sharing

App UOC

Es veu una mica estirat pel que recomanem visualitzar-lo al drive

https://cv.uoc.edu/app/blogaula212/212_m2_951_01_408662/2022/04/12/video-practica-1-tipologia-i-cicle-de-vida-de-les-dades-victor-baena-arnau-muns/

12. Taula de contribucions

Contribucions	Signatura
Investigació prèvia	**A.M.O | V.B.C**
Redacció de les respostes	**A.M.O | V.B.C**
Desenvolupament del codi	**A.M.O | V.B.C**

BIBLIOGRAFIA

(1) **https://www.idealista.com/ayuda/articulos/terminos-y-condiciones-generales-de-idealista/**