Àlex Penina Aguilera

Tipologia i cicle de vida de les dades

Pràctica 1 (35% nota final)

1) Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

La crisis derivada del Covid-19 ha canviat el mercat de l'habitatge en factors com el preu, les preferències en la recerca o la creixent necessitat de tenir un espai exterior per a suportar situacions com un confinament. Aquests canvis també s'han fet notables en el mercat del lloguer.

Abans de la pandèmia hi havia escassetat d'oferta d'immobles en lloguer, els preus eren elevats i els habitatges es llogaven ràpidament. Ara, per contra, s'està produint una alta rotació d'immobles i un increment de l'oferta, uns fets que estan comportant una reducció en els preus i un canvi en la recerca d'habitatges. Els arrendataris, donada la situació d'incertesa econòmica i laboral, estan interessats en pisos més barats i triguen més temps a prendre decisions.

En aquesta pràctica, volem analitzar aquest canvi de tendència a la ciutat més poblada de Catalunya: Barcelona. Com a internet hi ha multitud de portals de referència en la publicació d'immobles, hem seleccionat el portal immobiliari amb el major nombre de publicacions d'habitatges en lloguer a la capital catalana. Així, el portal d'Habitaclia ofereix, en data 4 de novembre del 2020, un total de 13.245 habitatges en lloguer oferts a la ciutat de Barcelona.

2) Definir un títol pel dataset. Triar un títol que sigui descriptiu.

Habitatges_Lloguer_Barcelona_20201104.csv

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 triat).

El dataset inclou tots els anuncis d'habitatges en lloguer a la ciutat de Barcelona que hi ha al portal immobiliari d'Habitaclia, el dia 4 de novembre del 2020 (darrera execució). Per a cada un dels immobles, obtenim les seves principals característiques:

- Barri
- Metres quadrats
- Nombre d'habitacions
- Nombre de lavabos
- Preu
- Tipus d'immoble

4) Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment.

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

Les dades fan referència a tots els habitatges en lloguer que hi ha a Barcelona anunciats a la web d'Habitaclia, en data 4 de novembre del 2020.

Les dades s'han recollit utilitzant tècniques de *web scraping* mitjançant codi Python i un fitxer Notebook de JupyterLab, fent ús de diferents llibreries especialitzades, com ara *BeautifulSoup* o *Pandas*.

Abans de realitzar *web scraping*, s'ha comprovat que el fitxer "robots.txt" no presentava cap restricció que afectés a l'extracció de les dades. No obstant, per tal d'evitar banejos, s'ha decidit camuflar l'*User Agent* del script simulant ser un navegador de Mozilla.

En el moment de l'extracció, hi havia un total de 13.245 habitatges en lloguer publicats, dividits en 884 pàgines (15 publicacions per pàgina). Degut a això, ha estat necessari crear un bucle que recorregués totes les pàgines i, en cada una d'elles, mitjançant expressions regulars, s'han extret les variables d'interès i s'han emmagatzemat en una llista. Per últim, s'ha transformat la llista en un *DataFrame* per tal de poder extreure les dades en format CSV.

Finalment, el nostre dataset inclou, per a cada habitatge, els camps següents:

Variable	Tipus	Descripció
Barri	Caràcter	Barri de Barcelona on pertany l'immoble
Tipus_Immoble	Caràcter	Tipologia d'habitatge (pis, casa, xalet, etc.)
Preu	Numèric	Preu mensual exigible per a llogar l'immoble
Metres_Quadrats	Numèric	Nombre de metres quadrats que té l'immoble
Num_Lavabos	Numèric	Nombre de lavabos que té l'immoble
Num_Habitacions	Numèric	Nombre d'habitacions que té l'immoble

6) Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

El propietari del conjunt de dades extret és Habitaclia, un portal immobiliari digital amb més de 17 anys d'experiència en el sector i un dels més utilitzats en tot el país, el qual és propietat de l'empresa Adevinta. Aquest portal, a més de publicar els anuncis de compra, lloguer, lloguer

de temporada i traspàs de tot tipus d'immobles a tota Espanya i Andorra, també s'encarrega de posar en contacte als compradors amb els venedors, als arrendataris amb els arrendadors, etc. sense cobrar cap tipus de comissió per la intermediació.

Agraeixo a l'empresa Adevinta i a la infraestructura informàtica del seu portal immobiliari d'Habitaclia per facilitar la possibilitat d'extracció de les seves dades sense cap tipus de restriccions.

No he trobat projectes de *web scraping* anteriors sobre el portal d'Habitaclia. No obstant, sí que n'he trobat d'altres del mateix estil com, per exemple, del portal Idealista:

https://github.com/David-Carrasco/Scrapy-Idealista

7) Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Tal com s'ha comentat en el primer apartat, he cregut convenient realitzar aquest estudi amb l'objectiu de mostrar la realitat amb la que es troba la població a l'hora de buscar un habitatge de lloguer a Barcelona en un context tan difícil com l'actual, en què la crisi de la Covid-19 ha tingut un impacte notori en tots els àmbits de la societat, incloent l'accés a l'habitatge.

Amb aquestes dades hi ha la possibilitat de realitzar un estudi sobre el preu mitjà dels habitatges en lloguer anunciats, segmentant-lo per tipus d'habitatge, per barri o per metres quadrats, per exemple.

Així, es podria donar resposta a les preguntes següents, entre d'altres:

- Quin és el preu de lloguer mitjà dels habitatges a Barcelona?
- A quin barri són més barats?
- Quin és el preu mitjà per metre quadrat?
- Per quin tipus d'immoble hi ha més anuncis?
- Quants habitatges hi ha de lloguer actualment?
- A quin barri de Barcelona és més car el metre quadrat?
- Les dades concorden amb les que publica l'Idescat sobre el preu mitjà de l'habitatge a la ciutat de Barcelona?
- Quin seria el preu de lloguer estimat d'un immoble de 90m2 i 3 habitacions en el barri de Sant Gervasi?

8) Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

- **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**
- **Other (specified above)**
- **Unknown License**

Pel dataset resultant del projecte s'ha seleccionat la llicència "CC-BY-SA 4.0 License".

S'ha optat per aquesta llicència per tal que qualsevol altre pugui fer ús del material (tot reconeixent l'autoria i indicant els canvis que s'hagin realitzat) i compartir-lo i/o adaptar-lo com cregui convenient, però sense fins comercials. Les condicions d'ús d'aquesta web especifiquen clarament que es prohibeix l'ús de les dades amb finalitats comercials.

Per a aquest projecte s'ha tingut en compte els termes i condicions de la web d'Habitaclia, així com del seu fitxer "robots.txt".

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

El codi Python amb el qual s'ha realitzat *web scraping* i generat el dataset és el següent:

```
#Importem les llibreries necessàries:
from bs4 import BeautifulSoup
import requests
import pandas as pd

#Definim les variables:
URL_BASE = "https://www.habitaclia.com/alquiler-barcelona-"
#Camuflem l'user agent simulant ser un navegador de Mozilla:
headers = {'User-Agent': 'Mozilla/5.0'}
#Tenim un total de 884 pàgines d'anuncis d'habitatges en lloguer a
Barcelona (04/11/2020):
NUM_PAGS = 884 comptador = 0 llista=[] Preus=[]
#Indiquem que apareixen 15 immobles per a cada pàgina (comptant el
0 com a primer valor):
Immobles_Per_Pagina=14

#Controlem les restriccions del fitxer robots.txt:
response = requests.get('https://www.habitaclia.com/robots.txt')
robots = response.text
print("\nFitxer robots.txt de https://www.habitaclia.com")
print("Verificació de restriccions:\n")
print(robots)
print("\nRealitzant scraping")

#Creem un bucle per a recórrer les pàgines d'habitaclia:
for i in range(1, NUM_PAGS):

if i < 1:
url = URL_BASE
else:
url = "%s%d.htm" % (URL_BASE, i)
print(url)
req = requests.get(url,headers=headers)

#Comprovem que la petició retorna Status Code = 200:
StatusCode = req.status_code
if StatusCode == 200:

#Passem el contingut HTML de la web a un objecte BeautifulSoup():
html = BeautifulSoup(req.text, "html.parser")

#Obtenim tots els divs de les entrades:
panell = html.find(id="js-list")
entrades = panell.find_all(class_="list-item-content")
```

```
#Recorrem totes les entrades per a l'extracció de les dades d'inter
ès:
for entrada in entrades:

#Barri de Barcelona on es troba l'immoble:
Barri = entrada.find(class_ = 'list-item-location').get_text()

#Tipus d'immoble de l'anunci:
Tipus_Immoble = entrada.find(class_ = 'list-item-title').get_text()

#Creem un bucle per a obtenir el preu de l'immoble:
Tot_preus = panell.select(".list-item-content-second .font-2")
Preus = [p.get_text() for p in Tot_preus]

#Metres quadrats de l'immoble:
Metres_Quadrats = entrada.find(class_= 'list-item-feature').get_tex
t()

#Nombre de lavabos de l'immoble:
Num_Lavabos = entrada.find(class_= 'list-item-feature').get_text()

#Nombre d'habitacions de l'immoble:
Num_Habitacions = entrada.find(class_= 'list-item-feature').get_tex
t()

#S'afegeixen els elements extrets a una llista:
llista.append({
"Barri": Barri,
"Tipus_Immoble": Tipus_Immoble,
"Preu": Preus[comptador],
"Metres_Quadrats": Metres_Quadrats,
"Num_Lavabos": Num_Lavabos,
"Num_Habitacions": Num_Habitacions
})

#Convertim la llista en una DataFrame:
df=pd.DataFrame(llista)

#Extraiem el contingut necessari a través d'expressions regulars:
df["Num_Habitacions"]=df.Num_Habitacions.str.extract('.+(\d+) habit
aciones')
df['Barri']=df.Barri.str.extract('Barcelona - (.+)')
df["Tipus_Immoble"]=df.Tipus_Immoble.str.extract('Alquiler (.+?)\s'
)
df['Metres_Quadrats']=df.Metres_Quadrats.str.extract('(\d+)')
df['Num_Lavabos']=df.Num_Lavabos.str.extract('.+(\d+) baño')

#Creem un comptador per a recórrer els 15 immobles que té cada pàgi
na:
comptador += 1
if comptador > Immobles_Per_Pagina:
comptador = 0
else:
#Si Status Code != 200, retorna "Error 400":
break

#Generem un fitxer csv a partir del DataFrame creat, indicant-li la
codificació 'utf-16', que no tingui en compte el número #de les fil
es i que el fitxer té capçaleres:
```

```
df.to_csv('Habitatges_Lloguer_Barcelona_20201104.csv',
encoding='utf-16', index=False, header=True)
```

Aquest codi també es pot veure en el fitxer "Codi_Python_PRAC1.ipynb" del *github*, junt amb el resultat de la seva execució:

https://github.com/apenina/PRAC1_Tipologia/blob/main/Codi_Python_PRAC1.ipynb

10) Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

El dataset en format CSV es pot veure en el fitxer "Habitatges_Lloguer_Barcelona_20201104.csv" del *github*:

https://github.com/apenina/PRAC1_Tipologia/blob/main/Habitatges_Lloguer_Barcelona_20201104.csv

També està publicat a Zenodo:

https://zenodo.org/record/4247744 - .X6RIOlNKhQI

El DOI de la publicació és el següent: 10.5281/zenodo.4247744

Taula de contribucions del treball.

Contribucions	Signa
Recerca prèvia	Àlex Penina
Redacció de les respostes	Àlex Penina
Desenvolupament del codi	Àlex Penina

Com ja li vaig comentar al professor col·laborador de l'assignatura, a l'organitzar-me les tasques del màster per assignatures vaig tardar uns dies de més en assabentar-me que calia formar grups per a la realització de la pràctica i, degut a aquest fet, no vaig trobar cap altre alumne disponible per a formar parella. El professor no va posar-me objecció a que realitzés la pràctica individualment.