Tipologia i cicle de vida de les dades
Pràctica 1 - Abril de 2020

Alumnes:

- Marc Serra Suñol
- Javier Beltran Lou

Respostes a les preguntes

Pregunta 1

Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.
El context d'aquesta pràctica és el mercat d'habitatges de lloguer concretament a l'àrea de Barcelona. El lloc web triat és **habitaclia.com**. Els motius principals per triar aquesta web són:

1. És una de les webs especialitzades amb més informació respecte al tema d'interès.
2. A diferència d'altres webs estudiades a l'anàlisi prèvia, aquesta en particular no prohibeix via robots.txt explícitament la recopilació de dades per part de tercers i tampoc, segons les notres proves, imposa dificulats extraordinàries en el procés de scraping, com podrien ser la detecció automatica de bots i les corresponents proves CAPTCHA.

Pregunta 2

Definir un títol pel dataset. Triar un títol que sigui descriptiu
Habitatges en règim de lloguer a l'àrea de Barcelona

Pregunta 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).
Llistat d'habitatges en règim de lloguer als municipis de l'àrea de Barcelona (Barcelona i rodalies) per poder fer estudis comparatius segons les diferents variables recollides.

Pregunta 4

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

Pregunta 5

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

Els següents camps del dataset s'han obtingut com a resultat del scraping de pàgines d'ofertes a habitaclia.com. Habitaclia fa més de 15 anys que publica en la seva pàgina web anuncis del mercat de l'habitatge en les principals metropolis espanyoles. Cada registre del dataset correspon a una oferta en el site. La majoria dels camps s'han obtingut a partir del markup de la pàgina, fent ús principalment de les diferents seccions i de diferents tractaments de cadena. Alguns dels camps s'obtenen directament de elements "input" amb un identificador ben definit, utilitzant també el markup, a mode de fallback, per si no estiguessin disponibles en alguna oferta en particular. Finalment, algunes dades s'obtenen directament a partir de la pròpia URL de cada oferta.

- **Id:** Identificador de l'habitatge
- **URL:** URL de la oferta obtinguda
- **Tipus Oferta:** Indica si es tracta d'un habitatge en Venda o en Lloguer. En el nostre CSV trobem només 'lloguer', que és el principal objectiu de la pràctica. Però val a dir que l'scraper pot obtenir informació també per ofertes de compra.
- **Tipus immoble:** Pis, Casa, Àtic o Dúplex
- **Municipi:** per exemple, 'Barcelona'

- **Província:** pe exemple, 'Barcelona'
- **Zona:** Barri o zona dins d'un municipi on es situa l'habitatge. Per exemple, 'Dreta de l'Eixample'
- **Preu:** Preu, en euros, de l'habitatge en el moment d'obtenir les dades
- **Data:** Data en què es va obtenir la informació de l'habitatge
- **Superfície:** Superfície, en metres quadrats, de l'habitatge
- **#Habitacions:** Nombre d'habitacions de què l'habitatge disposa
- **#Banys:** Nombre de banys de què consta la vivenda
- **Antiguitat:** Any de construcció de la vivenda
- **Planta:** Número de planta (1ª, 2ª, ...)
- **Parking:** booleà indicant si la vivenda disposa o no de pàrking
- **Calefacció:** booleà indicant si la vivenda disposa o no de calefacció
- **Aire acondicionat:** booleà indicant si la vivenda disposa o no de aire acondicionat
- **Moblat:** booleà indicant si la vivenda disposa o no de mobles
- **Ascensor:** booleà indicant si la vivenda disposa o no d'ascensor
- **Jardí:** booleà indicant si la vivenda disposa o no de jardí
- **Eficiència energètica:** Classificació de la vivenda segons la seva eficiència energètica (A..G),
- **Classe emissions:** Classificació de la vivenda segons les emissions (A..G)

Cal tenir en compte que no tots els habitatges disposen de tots els camps d'informació. Per aquells camps dels quals no es disposa d'alguna dada el valor reportat és un valor buit o ''.

Pregunta 6

Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).
Les dades han sigut recolectades de la pàgina web habitaclia.com. Per a fer-ho s'han utilitzat tècniques de webscraping, utilitzant el llenguatge de programació Python, per tal d'extreure informació de pàgines html. S'ha utilitzat la llibreria BeautifulSoup per a parsejar les pàgines html.

Pregunta 7

Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.
El mercat immobiliari és un dels més importants en qualsevol economia i la seva evolució a nivell de preus pot oferir indicador sobre la salut de la mateixa. Conjunts de dades similars al present, que es centra en les ofertes de lloguer d'una determinada zona, poden servir d'ajuda a l'hora d'analitzar el seu estat.

Alguns del usos o estudis que es poden plantejar a partir del dataset:

1. Estudi de diferències de preu / m^2 entre municipis, barris, districtes o zones, antiguitat, prestacions, ...
2. Evolució preu / m^2 (en el context de vàries extraccions en diferents moments) classificant per municipi, antiguitat, etc. Això es podria fer obtenint datasets en diferents moments. El scraping contempla la dada 'data' per a tal fi.
3. Increment mig de preu en els habitatges que disposen de parking i/o altres criteris
4. Comparatives creuant de les dades obtingudes en el dataset amb dades que podrien venir d'altres datasets, per exemple salaris a les zones estudiades, amb agregacions per zona, municipi, demarcació, etc.

Pregunta 8

Llicència
El present dataset es publica sota la llicència Released Under CC BY-NC-SA 4.0 License. Els motius per triar aquest tipus de llicenciament han estat:

- S'ha de dir d'on han obtingut les dades i els canvis que se'ls hi han fet
- L'ús de les dades no pot ser comercial, perquè la competència tregui benefici de l'única web que permet fer web scraping sense grans dificultats
- Han de llicenciar les noves dades de la mateixa manera, és a dir que l'ús que se'n pugui acabar fent segueixi sense repercutir negativament en el negoci de habitaclia.com.

Pregunta 9

Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
El codi font en Python es pot descarregar des de la URL
https://github.com/gikajavi/home-scraper/tree/master/src

Pregunta 10

Dataset. Publicar el dataset en format CSV a Zenodo amb una xicoteta descripció.
Link al dataset:
https://zenodo.org/record/3748054

Taula de contribucions

Contribucions	Signa
Recerca prèvia	MSS, JBL
Redacció de les respostes	MSS, JBL
Desenvolupament del codi	MSS, JBL