Pràctica 1: Com podem capturar les dades de la web?

Tipologia i cicle de vida de les dades aula 1

Arnau Gironella

Alexandre Fló

Índex

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.

És sabut que la gent quan va de vacances moltes vegades decideix passar l'estada en un hotel. Però, recentment han sortit empreses donant un nou punt de vista als allotjaments de curta estada. És el que es coneix com a lloguer vacacional, pràctica que es porta fent des de fa molts anys, però recentment s'ha modernitzat i digitalitzat.

En el nostre cas ens hem enfocat en la empresa més important a l'hora de reservar allotjaments d'altres persones i aquesta és *Airbnb*, un portal en línia on es pot reservar estades i experiències a curt termini. En aquest portal web trobem tota aquesta informació relacionada amb l'habitatge que ofereixen els diferents amfitrions, entre altres.

Lloc web: https://www.airbnb.es/

Títol. Definir un títol que sigui descriptiu pel dataset.

Un títol adequat per aquest data set seria: "Habitatges disponibles a Barcelona de desembre 2022. Dades d'Airbnb".

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

El data set inclou la informació bàsica dels habitatges com pot ser preu per nit, valoració general, la quantitat d'habitacions o si té disponibilitat de cuina, així com la valoració en altres aspectes de l'allotjament. Les dades corresponen al més de desembre del 2022 de la regió de Barcelona, Catalunya.

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

Alojamiento entero. Anfitrión: Julio

2 viajeros · 1 dormitorio · 1 cama · 1 baño

Julio es Superanfitrión
Los Superanfitriones son anfitriones con experiencia y valoraciones excelentes que se esfuerzan al máximo por ofrecer estancias inolvidables a sus huéspedes.

Ubicación fantástica
El 100 % de los últimos huéspedes han valorado con 5 estrellas la ubicación.

Cancelación gratuita durante 48 horas.

¿Qué hay en este alojamiento?

- Cocina
- Wifi
- Aparcamiento gratuito en las instalaciones
- Bañera de hidromasaje
- Admite mascotas
- ~~Detector de monóxido de carbono~~
- ~~Detector de humo~~

53 € noche ★ 5,0 · 10 evaluaciones

LLEGADA	SALIDA
27/1/2023	1/2/2023

HUÉSPEDES: 1 viajero

Reservar

No se te cobrará nada aún

53 € x 5 noches	264 €
Gastos de limpieza	5 €
Comisión de servicio	46 €
Total	**315 €**

¡Qué suerte has tenido! El alojamiento de Julio en Airbnb suele tener todas las fechas reservadas.

Denunciar este anuncio

Limpieza	5,0	Veracidad	4,9
Comunicación	5,0	Ubicación	4,9
Llegada	4,9	Calidad	5,0

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

Per cada habitatge disponible tenim els següents camps:

- **title**: Títol identificador de l'habitatge.
- **rate:** Puntuació mitjana del habitatge del 0 al 5.
- **evaluations**: Nombre de persones que han avaluat l'habitatge.
- **superanfitrion:** Indica si l'amfitrió es particular o superamfitrió.
- **place:** Lloc on està ubicat l'habitatge: país i regió.
- **price:** Preu per nit
- **limpieza:** Puntuació mitjana de la netedat de l'habitatge del 0 al 5
- **veracidad**: Puntuació mitjana de la veracitat de l'habitatge del 0 al 5
- **comunicacion**: Puntuació mitjana de la comunicació amb l'amfitrió del 0 al 5.
- **ubicacion:** Puntuació mitjana de la ubicació de l'habitatge del 0 al 5.
- **llegada:** Puntuació mitjana de l'arribada a l'habitatge del 0 al 5.
- **calidad:** Puntuació mitjana de la qualitat de l'habitatge del 0 al 5.
- **WIFI:** Indica si l'habitatge té WiFi.
- **TV:** Indica si l'habitatge té televisió.
- **cocina:** Indica si l'habitatge té cuina.
- **dormitorios:** Nombre de dormitoris de l'habitatge.
- **camas:** Nombre de llits de l'habitatge.

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.

El propietari de les dades és òbviament l'empresa *Airbnb*. Hi ha diverses empreses que es dediquen a fer recopilar informació de *Airbnb*, com és l'exemple de Data *Rabbu*, que s'enfoca més en Estats Units. Llocs similars a l'anterior seira *Airbtics* [3] que ho enfoquen també a una forma de treure rendibilitat al lloguer vacacional amb diferents dades i estadístiques.

En quant a aspectes legals, hem revisat el robots.txt de la web *Airbnb* i només hem agafat dades que permetia l'aplicatiu, els altres tot i ser interessants els hem deixat de banda com poden ser localització o les normes de la casa. El *User-Agent* que s'ha utilitzat es el *Googlebot*, ja que és el per defecte de *Google* Chrome.

Disalows mes enfocats a lusuari com pagar etc i en quant a dades qüestions de fotos i normes de la casa

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.

El conjunt de dades és prou interessant ja que inclou dades d'habitatges de tot Barcelona. Amb aquestes dades es pot analitzar molt bé tant els preus com la qualitat dels diferents vivendes.

Una possible us d'aquest data set es la realització d'un anàlisis de mercat en el cas que una persona vulgui posar una propietat en lloguer vacacional a Barcelona per tal de entendre quines són les millors condicions que ha d'oferir per tenir una propietat competitiva en el mercat.

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

La llicència seleccionada per el data set resultant seria la CC BY-NC-SA 4.0 License. La raó per la qual creiem que hauria de ser No comercialitzable és per que tot i nosaltres som estudiants del Màster de Ciència de Dades potser no trobem el potencial real d'aquestes dades, algú amb més coneixement si que podria utilitzar-les per fer benefici econòmic.

Codi

https://github.com/PandaPandula/PRACTICA1.TCVD

El que fem és bàsicament introduir els paràmetres de busca que necessitem, que en aquest cas són de Barcelona el desembre del 2022. Un cop ens apareixen els diferents allotjaments, per cada enllaç recopilem la informació desitjada. Un cop agafats els primers 20 resultats, passem a la següent pàgina i repetim el procés. El nombre de pàgines es variable, tantes com l'usuari necessiti.

Algunes de les dificultats que ens hem trobat han sigut els *pop-ups* i, que alguns dies el personal de *Airbnb* canviava alguns dels noms de les variables i hem hagut de retocar codi.

Dataset

Inclòs en el GitHub mencionat anteriorment. El DOI seria aquest:

10.5281/zenodo.7348658

Video

El vídeo explicatiu de la pràctica el trobem en el següent enllaç:

https://drive.google.com/file/d/1AwoPWubibFMU699iwTGQ9uEc6Ky3A-Iq/view?usp=share_link

Contribucions	Signatura
Investigació prèvia	Arnau Gironella, Alexandre Fló
Redacció de les respostes	Arnau Gironella, Alexandre Fló
Desenvolupament del codi	Arnau Gironella, Alexandre Fló
Participació al vídeo	Arnau Gironella, Alexandre Fló

Bibliografia

- **Rabbu (2022). Airbnb Data Analysis**

https://data.rabbu.com/

- **Airbtics (2022). Short Term Rental Analytics for High Return Investments**

https://airbtics.com/

- **Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.**