Tipología y Ciclo de vida de los datos

PRA1: Web Scrapping

Autores
Rafael Jimenez Sarmentero
Jorge Marchán Gutiérrez

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

En nuestra primera reunión vía Google Meet, lo primero que nos cuestionamos fue sobre qué área o materia iba a versar nuestra práctica. La intención era sugerir algunas propuestas y optar finalmente por una de ellas.

Como uno de los miembros del equipo trabaja actualmente en el sector inmobiliario, la idea de enfocarnos en este ámbito surgió enseguida.

El primer portal que nos vino a la mente, como no podía ser de otra manera, fue Idealista.

Escrutamos su fichero robots.txt y no observamos que estuviera desaconsejada la extracción automática de datos. Sin embargo, una lectura de las «Condiciones generales» nos disuadió de tal intención.

De modo que nuestra reunión continuó con una búsqueda de portales inmobiliarios que nos permitieran hacer el *scrapping* de manera legal.

Así, tras algunas búsquedas infructuosas, llegamos a Fotocasa <https://www.fotocasa.es/es/>.

Una inspección detallada nos hizo reparar en que, a través de la página principal podíamos extraer los enlaces a las diferentes viviendas. Asimismo, observamos que cada vivienda ofrecía una página de detalle en la que podían obtenerse distintos campos que serían de interés para la construcción de un *dataset:* área de la vivienda, número de habitaciones, número de cuartos de baño, etc.

2. Título. Definir un título que sea descriptivo para el *dataset*.

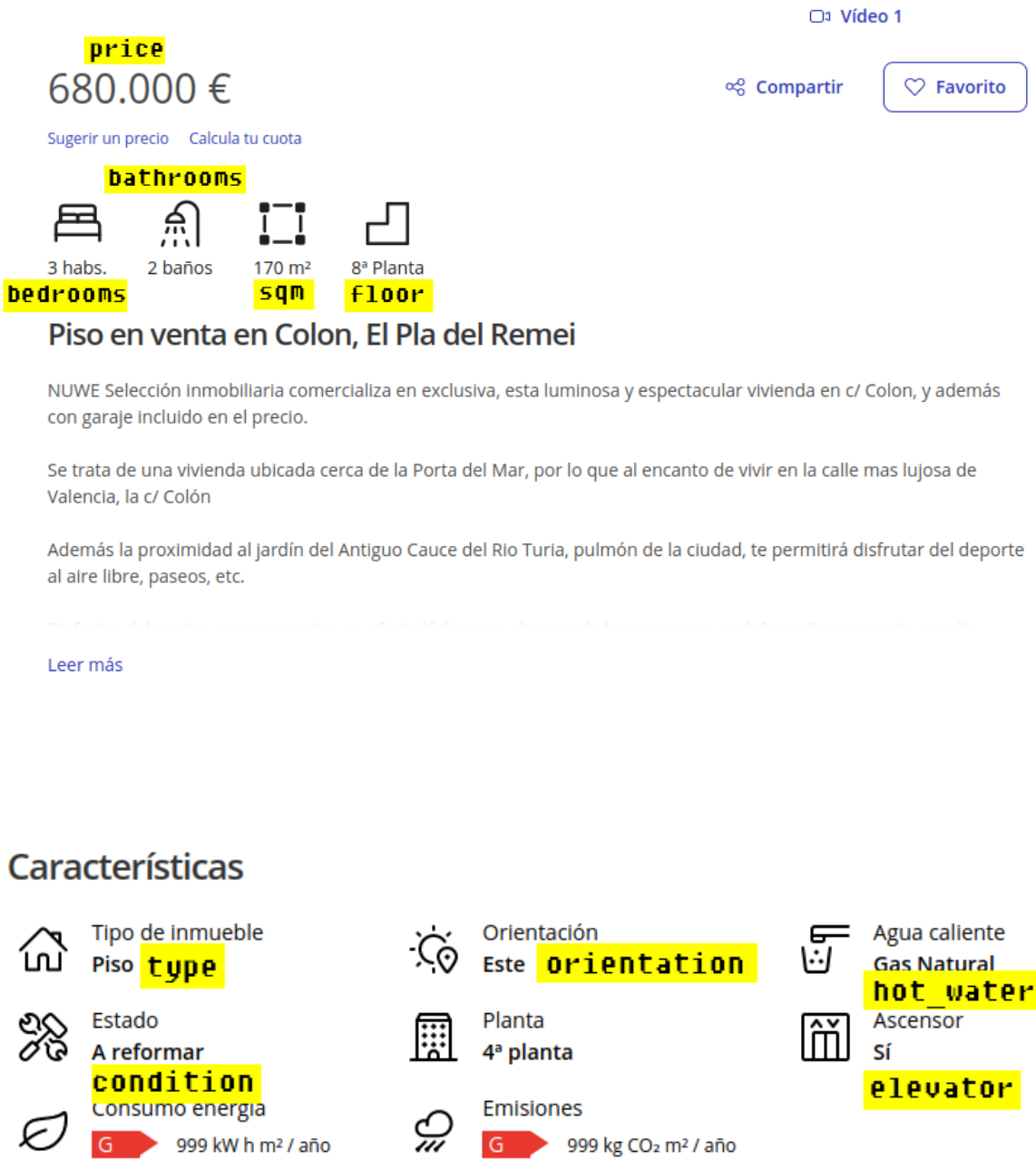
«madrid_real_state_pricing.»

3. Descripción del *dataset*. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El *dataset* corresponde con los datos de un conjunto de viviendas de Madrid, concretamente las tres primeras páginas de la búsqueda ya que según su robots.txt, Fotocasa solo permite búsquedas sobre las tres primeras páginas que conforman una búsqueda, sin embargo, el script se ha preparado para que se le pueda pasar el número de páginas que deseas obtener y el término de búsqueda, por lo tanto, podemos buscar para Madrid o para cualquier provincia o ciudad.

El *dataset* contiene información relativa al precio de la vivienda y sus características más básicas y que según nuestro criterio más pueden influir en este precio.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el *dataset* visualmente y el proyecto elegido.



Vídeo 1

price
680.000 €

Compartir Favorito

Sugerir un precio Calcula tu cuota

bathrooms

3 habs. 2 baños 170 m² 8ª Planta

bedrooms sqm floor

Piso en venta en Colon, El Pla del Remei

NUWE Selección inmobiliaria comercializa en exclusiva, esta luminosa y espectacular vivienda en c/ Colon, y además con garaje incluido en el precio.

Se trata de una vivienda ubicada cerca de la Porta del Mar, por lo que al encanto de vivir en la calle mas lujosa de Valencia, la c/ Colón

Además la proximidad al jardín del Antiguo Cauce del Rio Turia, pulmón de la ciudad, te permitirá disfrutar del deporte al aire libre, paseos, etc.

Leer más

Características

Tipo de inmueble
Piso type

Orientación
Este orientation

Agua caliente
Gas Natural
hot_water

Estado
A reformar
condition

Planta
4ª planta

Ascensor
Sí
elevator

Consumo energía
G 999 kW h m² / año

Emisiones
G 999 kg CO_2 m² / año

Ver etiqueta calificación energética

Gres Cerámica Cocina Office Patio Balcón Puerta Blindada

5. Contenido. Explicar los campos que incluye el *dataset*, el periodo de tiempo de los datos y como se han recogido.

Los campos que conforman el *dataset* extraído son los siguientes:

- **price:** Precio de la vivienda.
- **bedrooms:** Número de dormitorios.
- **bathrooms:** Número de cuartos de baño.
- **sqm:** Metros cuadrados de la vivienda.
- **floor:** Número de planta en la que se halla la vivienda.
- **type:** Tipo de la vivienda (piso, chalet, etc.).
- **orientation:** Orientación de la vivienda (norte, sur, este, oeste).
- **condition:** Estado de la vivienda (a reformar, etc.).
- **hot_water:** Forma de suministro del agua caliente.
- **elevator:** Si la vivienda dispone o no de ascensor.

En lo referente al periodo de tiempo de los datos es marzo de 2022, es decir, en el momento que se lanza el *script* el periodo de los datos con el cual se genera el *dataset* es “momento actual”.

En cuanto a la forma de recolección es navegando por los distintos resultados de las tres primeras páginas de búsqueda y accediendo a las distintas fichas de alojamiento para a continuación extraer los campos que hemos mencionado anteriormente, todo esto de manera automática, indicándole término de búsqueda y número de páginas.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para estar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El propietario del conjunto de datos es **fotocasa.es**, se trata de un portal inmobiliario dedicado a la compraventa y el alquiler de viviendas en España, hemos considerado que es interesante el web *scrapping* de este portal porque tiene suficientes datos y podríamos considerarlo un referente dentro del mercado de compraventa de inmuebles en nuestro país.

En cuanto a análisis, este portal tiene un blog y un *podcast* donde entre otras cosas hablan de los datos que poseen y de como los explotan, por lo tanto, tienen datos suficientes como para poder realizar análisis sobre ellos y llegar a conclusiones interesantes, como por ejemplo predecir datos en relación al mercado inmobiliario.

Para la realización del web *scrapping* se han respetado lo máximo posible los principios éticos y legales, en un primer lugar, se inspeccionó el contenido del fichero robots.txt en el que podemos ver como tienen una *blacklist* de algunos User-Agent a los cuales no permiten hacer *scrapping* de nada, pensamos que el motivo es que han detectado un abuso por su parte.

Para evitar esto, nosotros hemos instanciado tanto *Selenium* como *BeautifulSoup4* simulando que somos un navegador Chrome normal, haciéndonos pasar por un usuario de la aplicación. Por lo demás, podemos ver como en el fichero robots.txt permiten el *scrapping* de cualquier ficha de alojamiento, sin embargo, en lo relativo a la búsqueda solo permiten de las tres primeras páginas de cada búsqueda. Por este motivo, nuestro *scrapper* por defecto hace las tres primeras páginas de resultados, pero se puede adaptar para hacer más, o menos, así como jugar con el termino de búsqueda de forma que podemos generar varios *datasets* a partir de las tres primeras páginas de diferentes términos de búsqueda y concatenarlos para generar un *dataset* más grande.

Por último, de acuerdo a los principios éticos y legales, revisamos sus términos y condiciones y política de privacidad donde no encontramos ninguna referencia a este tipo de prácticas, por lo que legalmente la realización de la práctica es posible, sin embargo pese a no especificar ningún tipo de restricción, consideramos que la realización del *scrapping* de manera masiva puede resultar en un abuso e incluso podríamos llegar a degradar su servicio, perjudicando a otros usuarios o a la propia plataforma, por ello, realizamos el proceso de extracción de datos de manera secuencial, página a página, nunca en paralelo, asegurando de esta manera que no impacte en el rendimiento de la plataforma.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Consideramos interesante este conjunto de datos porque los inmuebles siempre es un sector de interés, ya sea para la inversión como para encontrar un lugar donde vivir acorde a las necesidades y capacidades de cada individuo. Es por este motivo que un *dataset* como el generado puede ser de gran utilidad, podría ayudar por una parte tanto a los particulares como a los profesionales, para hacerse una idea general del mercado, sin necesidad de realizar las búsquedas a mano y anotar los datos, permitiéndoles conocer el mercado y actuar en consecuencia y a otros sectores como analistas de datos que sobre el *dataset* pueden realizar todo tipo de análisis y predicciones.

8. Licencia. Seleccionar una de estas licencias para el *dataset* resultante y justificar el motivo de su selección:

- 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.

Por lo encontrado en sus políticas de datos y privacidad, fotocasa no prohíbe la extracción de datos, pero si su explotación comercial, es por este motivo por el cual, la licencia de las sugeridas que más encaja es **Released Under CC BY-NC-SA 4.0 License**.

9. Código. Adjuntar en el repositorio Git el código con el que se ha generado el *dataset*, preferiblemente en Python o, alternativamente, en R.

El código de la aplicación lo podéis encontrar en github donde además del código hay un fichero README.md que explica como ponerlo en marcha. Se ha realizado en Python utilizando Selenium y BeautifulSoup4.

10. Dataset. Publicar el *dataset* obtenido en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.z

En el repositorio se puede encontrar un fichero data.csv que contiene un ejemplo de un *dataset* generado, de viviendas de Madrid. Además, se ha subido el fichero generado a Zenodo, está disponible en esta URL con el siguiente DOI

11. Tabla de contribuciones

Contribuciones	Firma
Investigación previa	RJS, JMG
Redacción de las respuestas	RJS, JMG
Desarrollo del código	RJS, JMG