Práctica 1: ¿Cómo podemos capturar los datos de la web?

Adrián Valls Carbó, Javier Herrero Martín

Contexto

Con el objetivo de realizar un caso práctico sobre el uso de herramientas de extracción de datos automatizadas, hemos creído interesante hacer un estudio sobre los precios de la vivienda de alquiler en el parque inmobiliario español. Para ello, hemos elegido dos de los portales inmobiliarios más utilizados, Idealista (https://www.idealista.com/) y Fotocasa (https://www.fotocasa.es/es/)

Título del dataset

Datos de vivienda en alquiler en España

Descripción del dataset

Cada uno de los archivos contiene información sobre las diferentes viviendas de alquiler en España a día 22/11/2022, obtenido a partir de los portales inmobiliarios de idealista y fotocasa. Estos portales son las principales webs de búsqueda en castellano dentro del estado español y contienen información útil tanto para los agentes inmobiliarios como para los arrendadores sobre los diferentes inmuebles disponibles en una determinada ciudad. Hemos decidido extraer información de las ciudades de Madrid, Barcelona, Valencia, Sevilla, Bilbao y Tenerife, por ser las poblaciones con más habitantes y más oferta dentro de España.

Contenido

Se han obtenido diferentes sets de datos en función del portal del que se han extraído. En el caso de Fotocasa, se ha llamado ciudad_fotocasa.csv, donde ciudad es el nombre de la ciudad en la que se han buscado los datos, y se compone de las siguientes variables:

- **Ciudad**: Nombre de la ciudad en la que está situado el piso (str)
- **Precio**: Precio del piso en € / mes (int)
- **Superficie**: Superficie del piso en m^2 (int)
- **Habitaciones**: Número de habitaciones del inmueble (int)
- **Lavabos**: Número de baños del inmueble (int)
- **Planta**: Planta en la que se encuentra situada el inmueble (str)

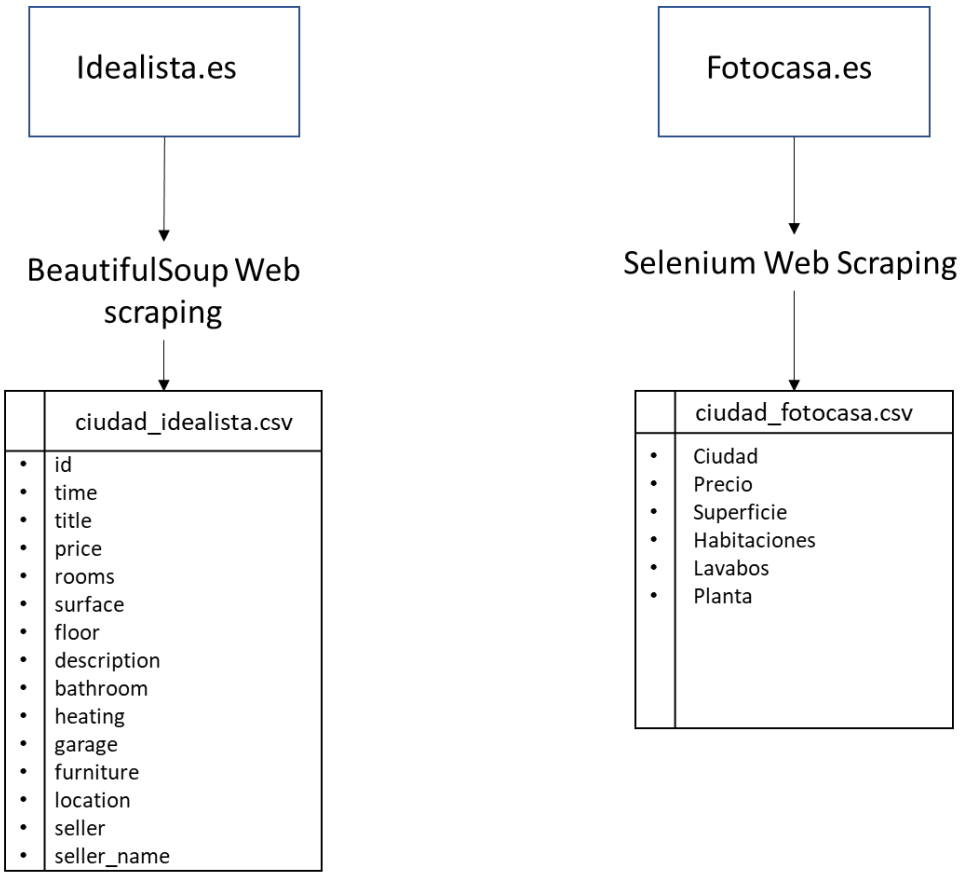
Los datos han sido extraídos en el día 22/11/2022 (estas páginas actualizan constantemente su feed)

Para el caso de Idealista, se ha llamado ciudad_idealista.csv, donde ciudad es el nombre de la ciudad en la que se han buscado los datos, y se compone de las siguientes variables:

- **id:** Referencia asociada al anuncio del piso (int)
- **time:** Fecha y hora en la que se ha obtenido la información (datetime)
- **title:** Nombre asociado al anuncio (str)
- **price:** Precio del piso en € / mes (int)
- **rooms:** Número de habitaciones del inmueble (int)
- **surface:** Superficie del piso en m^2 (int)
- **floor:** Planta en la que se encuentra situada el inmueble (str)
- **description:** Descripción asociada al anuncio (str)
- **bathroom:** Número de baños del inmueble (int)
- **heating:** Tipo de calefacción del inmueble, si tiene (str)
- **garage:** Plaza de garaje asociada al inmueble, si tiene (str)
- **furniture:** Muebles con lo que cuenta el piso (str)
- **location:** Localización del inmueble (str)
- **seller:** Tipo de vendedor del inmueble (str)
- **seller_name:** Nombre del vendedor (str)

Los datos han sido extraídos en el día 22/11/202, tal como puede verse en la variable time de cada uno de ellos (estas páginas actualizan constantemente su feed)

Diagrama



Propietario

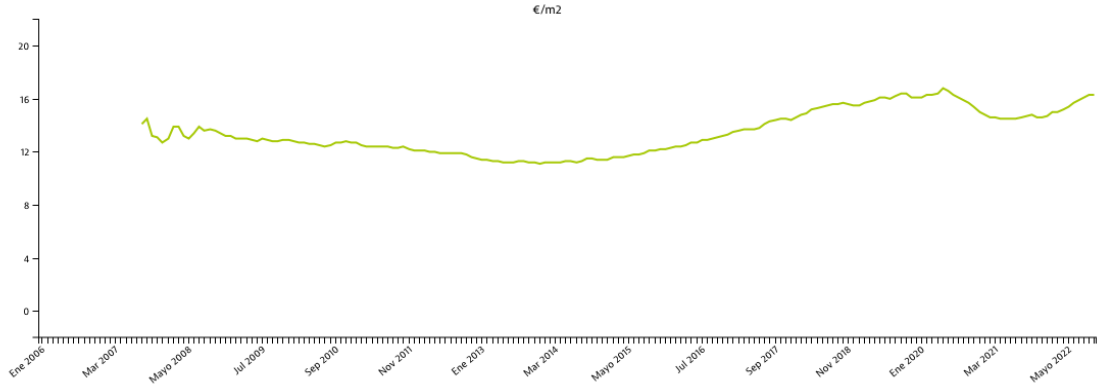
Los propietarios de los portales web de dónde se han obtenido los datos, así como su información relacionada, son:

- Fotocasa.es:
    - Fotocasa es el portal inmobiliario líder de España y cuenta con 1,5 millones de inmuebles de segunda mano, promociones de obra nueva y viviendas de alquiler. Fue fundado en 1999 y lleva más de veinte años ayudando a encontrar un hogar a millones de españoles. Fotocasa pertenece a Adevinta, una compañía líder en marketplaces digitales y una de las principales empresas del sector tecnológico del país, con más de 18 millones de usuarios al mes en sus plataformas de los sectores inmobiliario (Fotocasa y habitaclia), empleo (Infojobs.net), motor (coches.net y motos.net) y compraventa de artículos de segunda mano (Milanuncios).
- Idealista.es :
    - idealista (Idealista, S.A.U.) es la empresa que gestiona esta Web y Apps. La sede está en Madrid, en Plaza de las Cortes. Inscritos en el Registro Mercantil de Madrid y con NIF A-82505660. Facilitan un espacio en el que publicar, o buscar, anuncios de venta o alquiler de inmuebles, sea un piso, una habitación en piso compartido o un garaje. También ofrecen otros servicios relacionados con el sector inmobiliario, como valoraciones de inmuebles o el servicio de certificación energética.

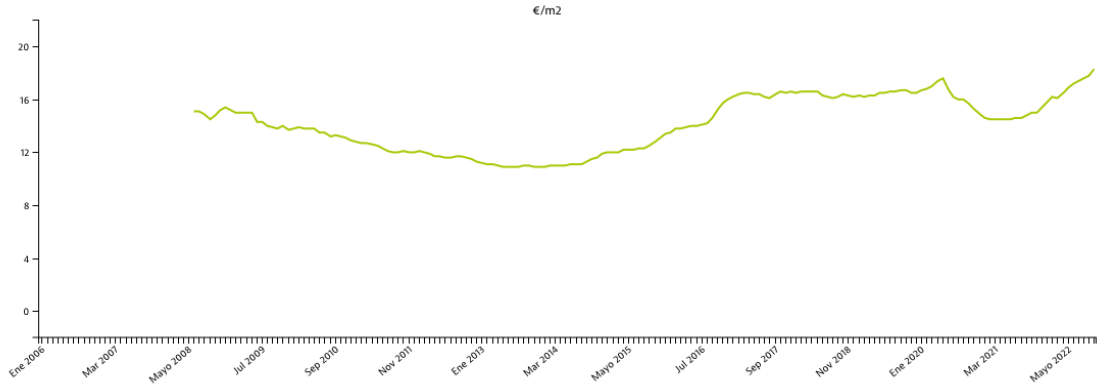
En cuanto a análisis similares, se han realizado análisis de la vivienda de alquiler desde diferentes ámbitos:

-Fotocasa Research (https://research.fotocasa.es/): Esta división de la compañía fotocasa, se define según su página web como un área de análisis enmarcada dentro del departamento de comunicación "con el objetivo de ofrecer información relevante del sector inmobiliario en España que sirva de ayuda tanto a los profesionales a conocer mejor el entorno en el que ejercen su profesión como a los usuarios a tomar la mejor decisión tanto si quieren vender, comprar o alquilar". Desde este grupo se realiza un análisis del precio medio de la vivienda, así como del precio por metro cuadrado, que está disponible para todo el territorio español en la web https://www.fotocasa.es/indice-precio-vivienda/

- Idealista (https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/): desde la sección de prensa del portal se realiza un análisis del precio por metro cuadrado en cada una de las provincias del estado, analizando mes a mes la variación en el precio del alquiler. En la figura que se muestra a continuación se aprecian los cambios en el precio del alquiler en los últimos 5 años en la ciudad de Madrid, apreciándose la tendencia al encarecimiento del suelo.



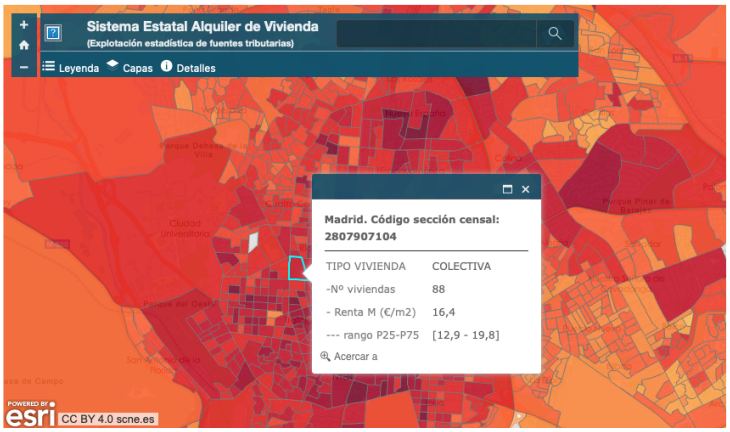
Precios de la vivienda en Madrid (€/m2)



Precios de la vivienda en Barcelona (€/m2)

- Gobierno de Cataluña: con motivo de la ley 11/2020 por la regulación del precio del alquiler, sobre la que posteriormente se declaró un recurso de inconstitucionalidad en el año 2022, se obtuvo un índice de los precios de alquiler, aunque no figura el modo de obtención de esta información, probablemente proceda de la información obtenida en el INCASOL, el organismo oficial de depósito de las fianzas de los alquileres.

- Gobierno de España: con motivo de garantizar la transparencia del mercado de la vivienda, el ministerio de transportes, movilidad y agenda urbana, publica a nivel nacional un registro de los índices de precios de la vivienda por cada uno de los municipios del estado. Estos datos proceden de la explotación de información tributaria, no disponibles de forma abierta e informan del precio por metro cuadrado en cada una de las secciones censales.



Inspiración

Si bien ninguna de las dos páginas permite de manera abierta la extracción de datos de sus portales, fotocasa es la más accesible sin el uso de APIs propias. Como ya hemos visto en el apartado anterior, también es la que ofrece información a través de sus filtros para conocer cómo varía el precio de alquiler o compra según diferentes filtros aplicados, como zonas o barrios, y desglosa dicho precio según ciertas variables, como precio/m^2 , precio/habitación o precio con terraza.

Idealista en cambio es bastante más hermética con sus datos, por lo que la idea es obtener datos de este último portal y realizar con ellos un análisis similar al que realiza Fotocasa. Además, realizaremos la extracción de ciertas variables básicas de Fotocasa mediante Selenium, para probar una técnica diferente, y así tener cierta comparación con los análisis originales y poder comprobar la representatividad de nuestro dataset.

Licencia

Debido a la política de los portales web de ambas compañías, se ha elegido licenciar los datos bajo el estándar Creative Commons Attribution Share Alike 4.0 International, de manera que los datos no concedan ningún derecho de patente a los que los utilicen, así como asegurar que otras licencias que utilicen estos datos tengan que tener una licencia similar.

Código

Los datos del código para explotar la información están disponibles en https://github.com/adrianvallsc/webscraping_housing de forma abierta.

Entre las principales dificultades a la hora de realizar la obtención de datos desde la web idealista se encuentra la ausencia del archivo robots.txt, además de las diferentes trabas que impone la web para realizar scraping. Esto hace que en numerosas ocasiones, cuando se realizan múltiples peticiones al servidor, se bloquee al usuario requiriendo la resolución de un

captcha. Aunque en el código se han intentado diversas formas para espaciar las peticiones y así no saturar los servidores, la web continúa realizándolo. Por este motivo, ha sido necesario obtener la cookie guardada tras el captcha (disponible en el archivo source.variables.py, dentro del diccionario cookie["datadome"]). En caso de volver a realizar un bloqueo, sería necesario obtener de nuevo esta cookie desde el captcha para poder resolver el problema. Otro problema importante es el tiempo de ejecución, que se ha conseguido mejorar limitando el número de páginas en general a 2, lo que reduce el tiempo de respuesta.

En lo que respecta a trabajar con Selenium, se debe en primer lugar instalar el paquete y elegir el navegador a utilizar. Ya que se simula el uso natural de la página web, se debe ir examinando esta para encontrar los botones que se deben clickar en cada caso y dónde poder escribir si es necesario. Por otro lado, para acceder a los datos se debe scrollear la página, lo que hace que la obtención de los mismos sea más lenta que por otros métodos, pero permitiendo una mejor funcionalidad y granularidad en aquellas páginas de scroll infinito, como reddit o twitter. Finalmente, se deben realizar ciertas pausas en la navegación, tanto para darle tiempo a la página a cargarse (recordemos que se abre un navegador real) como para no saturar la página con peticiones.

Dataset

El dataset obtenido de la extracción de las ciudades de Madrid, Barcelona, Valencia, Sevilla, Bilbao y Tenerife está disponible en Zenodo (https://doi.org/10.5281/zenodo.7348598)

Vídeo

Siguiendo el enlace que aparece a continuación, se puede encontrar el vídeo explicativo del proyecto y el código

https://drive.google.com/drive/folders/1RDfsWIwBx6zsNnqnI6BptSnP_-_IsSvH?usp=sharing

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

● Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

● Masip, D. (2019). El lenguaje Python. Editorial UOC.

● Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

● Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.

● Tutorial de GitHub https://guides.github.com/activities/hello-world.

Contribuciones	Firma
Investigación Previa	AVC, JHM
Redacción de las respuestas	AVC, JHM
Desarrollo del código	AVC, JHM
Participación en el vídeo	AVC, JHM