Universitat Oberta de Catalunya

Alba Gómez Varela · agomezvarela@uoc.edu
Patricia Lázaro Tello · plazarotello@uoc.edu

Web scraping del mercado inmobiliario

Alba Gómez Varela
Patricia Lázaro Tello

Título: Houses for sale in the Salamanca and Villaverde district of Madrid in April 2022

Licencia: CC BY-NC-SA 4.0

DOI:

Número de campos: numero

Número de registros: numero

Fecha de extracción: primera semana de abril de 2022

1. Contexto

La **Ley por el Derecho a la Vivienda** se aprobó en España por decreto el 1 de febrero de 2022 y, aunque esta normativa no regula los pisos en venta ya construidos en la actualidad de forma directa, sí que podría afectar a su precio en el futuro. Por ejemplo, si se limita el precio del alquiler, se puede dar el caso de que los inversores dejen de estar interesados en estos bienes y, por tanto, el precio final de venta variaría al cambiar la demanda. Por otro lado, el 30 por ciento de las nuevas promociones se reservará a Viviendas de Protección Oficial (VPO), lo que también puede alterar el precio de la vivienda libre.

Además de cambios en la normativa del mercado inmobiliario, la **invasión de Rusia a Ucrania** ha pillado por sorpresa a todos los sectores de la economía, disparando la **inflación** en España hasta un 9,8 por ciento, según indicador adelantado del IPC para el mes de marzo publicado por el Instituto Nacional de Estadística. Del mismo modo, el Banco de España ha recortado sus **previsiones de crecimiento** para el país y ha doblado su previsión anual de inflación al 7,5 por ciento, según los datos publicados el 5 de abril de 2022.

En este contexto de cambios económicos y normativos, el mercado inmobiliario en España se ha caracterizado desde hace décadas por las **diferencias en el precio** de la vivienda, tanto según las zonas del territorio nacional como dentro de las propias localidades.

Teniendo presente esta situación, y no existiendo un registro público de viviendas en venta actualizado, se ha decidido obtener la información de las viviendas en venta de

los portales inmobiliarios de **Idealista** y **Fotocasa**. Esta decisión se fundamenta en que son los **más visitados** en España, según los datos de febrero de *similarweb*, que publica las métricas oficiales de los sitios web de todo el mundo. Además, esta clasificación de visitas se mantiene en los últimos años, por lo que no se prevé que varíe en un corto periodo de tiempo. Asimismo, es interesante que estos portales trabajan tanto con **inmobiliarias** como con **particulares** que desean vender inmuebles, por lo que proporcionan una información amplia sobre el tema que nos ocupa en este proyecto.

2. Título del dataset

Debido a la gran cantidad de conjuntos de datos existentes sobre la temática seleccionada, para el dataset resultante se ha optado por un título **muy descriptivo** a la par que **concreto**, de modo que solo con él se pueda saber el contenido del mismo. Además, debido a la volatilidad del mercado inmobiliario, se ha decidido incluir el mes de extracción de los datos. Así, el título es el siguiente:

Houses for sale in the Salamanca and Villaverde district of Madrid in April 2022

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.

4. Representación gráfica

Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

5. Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido

Alba Gómez Varela · agomezvarela@uoc.edu
Patricia Lázaro Tello · plazarotello@uoc.edu

6. Agradecimientos

A continuación se abordará el apartado de agradecimientos divido en tres bloques debido a la complejidad del proyecto planteado y la necesidad de diferenciar claramente entre ellos.

Propietarios de los datos

Como se ha indicado en otros apartados nos encontramos ante un proyecto en el que se integran **dos fuentes de datos** diferentes, con sus respectivos propietarios de los datos.

Por un lado, sobre el portal web **Idealista**, cuyo fundador y CEO es Jesús Encinar, es importante destacar que la empresa está registrada como una Sociedad Anónima Unipersonal (Idealista S.A.U) fundada en el año 2000, aunque en la actualidad está participada mayoritariamente por los fondos gestionados por EQT. En estos momentos opera además en Portugal e Italia, países en los que no nos centramos en este proyecto, pero cuya estructura de la web es idéntica, con todo lo que esto implica.

Explorando el dominio mediante *whois* de python, se obtienen otros datos del mismo, como que su fecha de creación fue el 21 de agosto de 1999 y otra información irrelevante en este punto.

Por otro lado, según explican ellos mismos y se puede verificar en gran medida en el desarrollo del código, el **desarrollo de software** de idealista usa como base el sistema de arquitectura LEMP (Linux, Nginx, MySql, PHP) además de otras herramientas, como bases de datos (Oracle y SQLite), motores de búsqueda fulltext (Sphinxsearch y Solr), lenguajes de programación PHP en sus versiones 5.6 y 7.1 ó JavaScript y Laravel y Lumen como frameworks para el desarrollo de aplicaciones.

En cuanto al portal web **Fotocasa**, pertenece a Adevinta Spain S.L.U., empresa que también es propietaria del portal inmobiliario **Habitaclia** y se fundó en 1999. Al contrario que con Idealista, la estructura de los portales de los que Adevinta es propietaria es diferente en cada caso. Con Fotocasa, además, no se obtiene ningún tipo de información del dominio mediante *whois*.

Por otro lado, son menos transparentes en cuanto a cómo está **desarrollado** el portal, aunque sí que se han detectado el uso de bases de datos, motores de búsquedas y lenguajes de programación PHP y JavaScript.

Trabajos anteriores

Idealista y Fotocasa han sido objeto de análisis en varias ocasiones, y los que se han hecho públicos proceden sobre todo del **ámbito académico**, debido a la dificultad que conlleva el intentar someter a estos portales a un proceso de *scraping*. La razón de esto

reside en que el corazón de su negocio son los datos y cualquiera que se haga con ellos puede obtener una ventaja competitiva indeseable para ambos portales. Es por este motivo por el que sus equipos de desarrolladores son especialmente conocidos por intentar poner trabas a estas prácticas que, como se verá en el siguiente apartado, de manera explícita ellos prohíben realizar.

En el contexto del Máster en Inteligencia Artificial aplicada a los Mercados Financieros (mIA-X) del Instituto BME, David Carrasco Cuñado desarrolló un proyecto para *scrapear* **Idealista**, cuyo código se puede encontrar en **este repositorio** de GitHub y tiene como última actualización octubre de 2020. Este repositorio es el más completo que se ha encontrado que se aproxime a los objetivos del nuestro y únicamente permite registrar la información de cada uno de los inmuebles que aparecen en la **primera página de una URL** dada con una ubicación en concreto, extrayendo los datos únicamente de esta dirección sin entrar a ver los detalles en las páginas correspondientes de cada una de las viviendas.

El tutorial **más completo** es el ofrecido por el usuario de YouTube Code Monkey King, que se puede ver en **este vídeo**. El punto de entrada son códigos postales y enseña a descargar todos los datos tanto de las páginas de localización en las que se encuentran los listados de inmuebles como de las URL con cada uno de ellos. Por tanto, aunque no proporcione el código ejecutable de una forma directa, se trata de una pieza completa e interesante.

Por su parte, Miguel Ángel Gisbert publicó en junio de 2021 un tutorial sobre cómo *scrapear* los datos de una vivienda **dada la URL** de la misma, que se puede consultar en **este vídeo**. Al igual que el proyecto anterior, no realiza ningún análisis una vez obtenidos los datos.

En cuanto a Idealista, ya no destacan más ejemplos que merezcan una especial mención puesto que no se acercan, ni de lejos, a los objetivos de este proyecto. Por ejemplo, se ha encontrado **una alternativa** al *scraping* directo del portal mediante el análisis con R de la propia bandeja de entrada del **correo electrónico**, al que llega información de los pisos que cumplen los requisitos a los que el usuario se ha suscrito.

Centrándonos en **Fotocasa**, los compañeros Irene Fernández Molina y Héctor Hernández Membiela en el contexto académico de esta misma asignatura en abril de 2019 desarrollaron un proyecto de *web scraping* en el que este era uno de los dos sitios web del que realizaban la captura de datos. Como se puede observar en **su repositorio**, sin embargo, toda la información que se extrae se encuentra en las páginas en las que se muestra el **listado de inmuebles** y no se propone una extracción de la información que contiene cada una de las páginas de los mismos.

También en el ámbito académico, el Trabajo de Fin de Máster (TFM) de Álvaro Torrente Patiño del Máster de Ciberseguridad de la Universidade da Coruna también trabajaba sobre el *web scraping* de portales inmobiliarios, entre los que se encontraba Fotocasa según ha explicado él mismo en LinkedIn, pero el portal de la facultad en el que está colgado **dicho trabajo** ha estado caído durante todo el desarrollo de este proyecto, por lo que no se ha podido estudiar su trabajo previo.

Por otro lado, el usuario de YouTube Fpred publicó en octubre de 2020 un **pequeño tutorial** en el que explica cómo enfrentarse al *lazy loading* de la web de Fotocasa para obtener los precios de todas las viviendas que aparecen en cada página del listado cuando se busca una localización, como se puede ver en **este vídeo**, pero no profundiza en ningún otro aspecto adicional.

En relación a **análisis** en sí mismos, no se ha encontrado ninguno relevante ni con datos descargados de Idealista ni de Fotocasa más allá que los que ellos publican en sus respectivos portales web. Así, cabe destacar, por último, que es probable que existan **empresas** que hayan realizado proyectos para obtener la información de estos dos portales web, pero no se han hecho públicos y, por tanto, no se han podido tener en cuenta para esta fase del proyecto.

Pasos para actuar con principios éticos y legales

Este es uno de los apartados **más complejos** del proyecto y que hemos tratado con especial sensibilidad.

Según se recoge en el fichero **robots.txt** de **Idealista**, están específicamente no permitidas algunas de las acciones contempladas en este proyecto, como la navegación a través de la **paginación** (”Disallow: /*/pagina-*.htm”) o el empleo de más de dos **filtros** (”Disallow: /venta-*,*,*,.º ”Disallow: /venta-*,”), cuando en este caso se están empleando tres: viviendas + Madrid + distrito.

Según se recoge en el fichero **robots.txt** de **Fotocasa**, por su parte, el portal es mucho menos restrictivo. Sin embargo, también está especificado que no se permite la navegación a través de la **paginación** a partir de la cuarta incluida (”Disallow: */l/4*”hasta ”Disallow: */l/39*”), condición necesaria para el proyecto planteado. No obstante, para la búsqueda de las viviendas de cada distrito no se incumple ninguna restricción ya que, según la construcción de la URL no está especificada como tal. Por ejemplo, el archivo indica ”Disallow: /*filter=*”pero la URL de Villaverde es https://www.fotocasa.es/es/comprar/viviendas/madrid-capital/villaverde/l?sortType=scoring. Además, este tipo de *sortType* es el único permitido, por lo que tampoco entra en conflicto con las indicaciones del propietario.

Así, y teniendo en cuenta que estas restricciones son ”solo una sugerencia y **nunca una**

obligación", tal y como se señala en el documento 'Web scraping', además de tenerlas en cuenta para reducir las posibilidades de ser bloqueados, se decide que este proyecto tenga una finalidad **exclusivamente educativa**. Es interesante que trabajando ambos portales, las posibilidades de aprendizaje se disparan exponencialmente pero, sin embargo, no debemos olvidar la voluntad de los propietarios de los mismos. Por ello, en todo momento se intenta mantener el equilibrio y, sobre todo, **respeto** hacia Idealista y Fotocasa.

Agradecimiento final

Debido a todo lo expuesto en este apartado, agradecemos tanto a **Idealista** como a **Fotocasa** el haber servido como soporte para el proceso de aprendizaje que hemos desarrollado durante este proyecto de captura de datos, los cuales serán empleados con un fin exclusivamente académico. Además, hacemos extensible este agradecimiento a las personas mencionadas en **trabajos anteriores** por haber hecho público el conocimiento que han adquirido y que ha servido de punto de partida para este proyecto o para solventar obstáculos que nos hemos encontrado.

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 agradecimientos

Beautiful soup

8. Licencia

La licencia seleccionada es **Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)**, debido a que es la que más se ajusta a las necesidades de este proyecto y las preferencias de las integrantes del equipo de trabajo por los siguientes motivos:

● **Atribución**. Aunque el proyecto se puede compartir y adaptar, se debe atribuir a las creadoras del mismo, con un enlace a la licencia e indicando los cambios que se han realizado. El objetivo es que se reconozca el trabajo realizado en esta fase y las modificaciones respecto al original sin que, en ningún caso, se sugiera que este uso tenga el apoyo de las licenciantes.

● **NoComercial**. Este punto ha sido el más relevante a la hora de seleccionar la licencia

porque no está autorizado el *scrapeo* de los portales web de Idealista y Fotocasa. Por esta razón, este proyecto se ha desarrollado exclusivamente con fines académicos y, por tanto, queda terminantemente prohibido que se emplee el material con propósitos comerciales.

● **CompartirIgual**. Por el mismo motivo que no se permite una explotación comercial ni del proyecto ni de las modificaciones que se pudieran trabajar sobre él, cualquier contribución en el que esté involucrado debe tener la licencia CC BY-NC-SA 4.0. De esta manera, se evitará un uso no deseado de cualquier transformación.

Como se puede observar, el mayor interés a la hora de seleccionar esta licencia, por tanto, es favorecer el **flujo de conocimiento** con fines exclusivamente académicos a la vez que preservar la **voluntad de los propietarios** de los datos originales.

9. Código

El **código** para el desarrollo de este proyecto se puede consultar al completo dentro de *house-scraper* en: https://github.com/plazarotello/web-scraping

10. Dataset

Blablablá

11. Vídeo

Blablablá

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	AGV, PLT
Desarrollo de código	AGV, PLT
Redacción de las respuestas	AGV, PLT

Referencias

[1] Jaime Buelta. *Web scraping*. Packt Publishing, sep. de 2018. URL: https://learning.oreilly.com/library/view/python-automation-cookbook/9781789133806/ (visitado 29-03-2022).

[2] Stacy Creasey. *Mimicking Human Activity Using Selenium And Python*. 25 de ago. de 2021. URL: https://binarydefense.com/mimicking-human-activity-using-selenium-and-python/ (visitado 31-03-2022).

[3] Richard Lawson. *Web scraping with Python*. Packt Publishing, oct. de 2015. URL: https://www.packtpub.com/product/web-scraping-with-python/9781782164364 (visitado 01-04-2022).

[4] Ryan Mitchell. *Web scraping with Python*. O'Reilly, jul. de 2015. URL: https://www.oreilly.com/library/view/web-scraping-with/9781491985564/ (visitado 01-04-2022).

[5] Oren Spiegel. *The Art of Not Getting Blocked: How I used Selenium and Python to Scrape Facebook, and Tiktok*. 12 de nov. de 2019. URL: https://medium.com/analytics-vidhya/the-art-of-not-getting-blocked-how-i-used-selenium-python-to-scrape-facebook-and-tiktok-fd6b31dbe85f (visitado 01-04-2022).

[6] Laia Subirats Maté. *Web scraping*. Universitat Oberta de Catalunya, 18 de nov. de 2018. URL: https://materials.campus.uoc.edu/daisy/Materials/PID_00256970/pdf/PID_00256970.pdf (visitado 01-04-2022).