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

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

DOI: 10.5281/zenodo.6418974

DOI 10.5281/zenodo.6423459

Extracción: 4-7 de abril de 2022

Número de registros: 3.545

Número de campos: 16

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 % 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 %, según indicador adelantado del IPC para el mes de marzo publicado por el Instituto Nacional de Estadística (INE). 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 %, según los datos publicados el 5 de abril de 2022.

Asimismo, cabe destacar que 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 de cambios económicos y normativos, 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 portales **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 de interés del proyecto.

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

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.

3. Descripción del *dataset*

El conjunto de datos obtenido a partir de los *scrapers* hace referencia a todos los pisos disponibles durante la primera semana de abril de 2022 para su compra en los distritos de Villaverde y Salamanca en la ciudad de Madrid en los sitios web de Idealista y Fotocasa.

Se trata por tanto de una *snapshot* del mercado inmobiliario en el distrito con el precio/m^2 más bajo (Villaverde) y más alto (Salamanca) en Madrid. Con esta selección de distritos se busca descubrir las características que poseen los pisos de cada distrito, y si existe alguna otra razón, además del distrito en que se encuentra cada vivienda, que justifique su precio.

Cabe destacar que el conjunto de datos resultante no ha pasado por un proceso de limpieza, por lo que pueden existir inconsistencias en algunos campos; por ejemplo, los registros correspondientes a Fotocasa presentan el piso con un único número, mientras que en Idealista se muestran siguiendo el formato `Planta {X}` (por ejemplo, “Planta 1ª interior sin ascensor”). Otro ejemplo es el precio de los inmuebles, que en algunos casos se encuentra como *string* con signos de puntuación.

El *dataset* está guardado con codificación **UTF-8**, y cada registro se puede identificar inequívocamente mediante el par `{id, source}`.

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

4. Representación gráfica

A continuación se muestra el flujo de este proyecto:

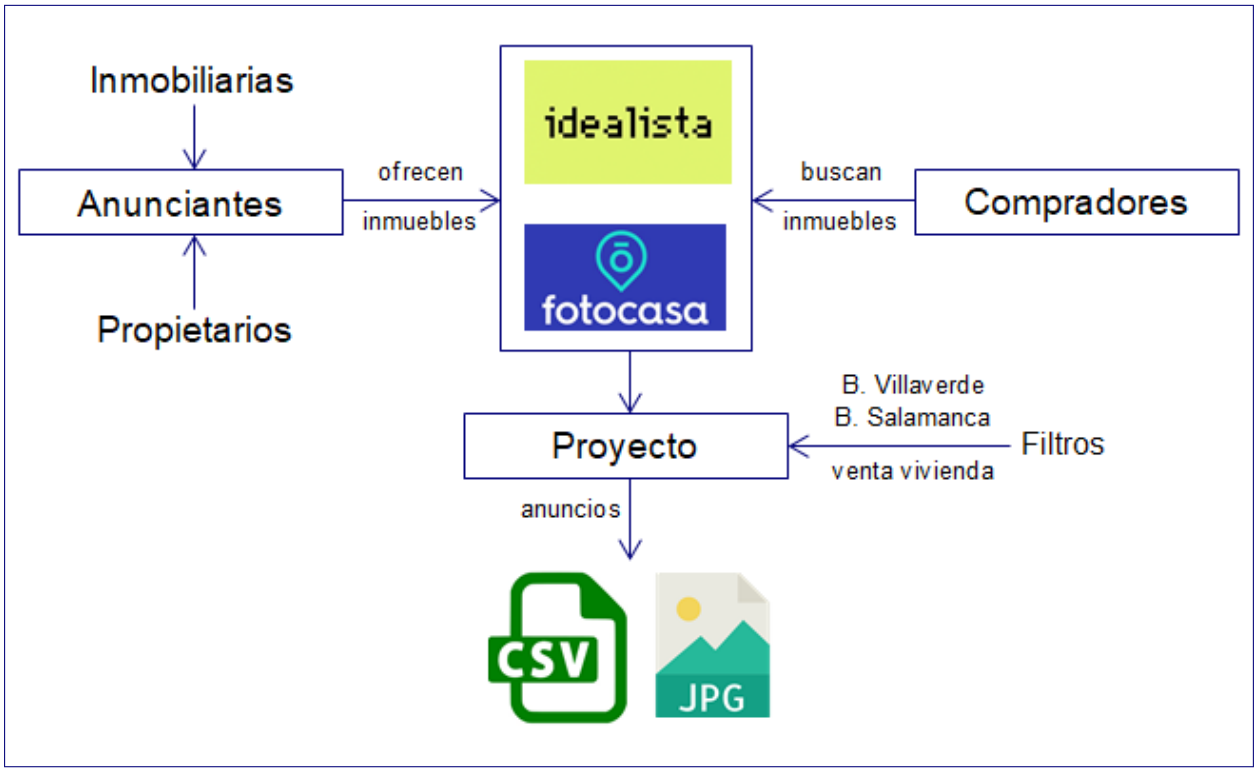


Figura 1: Representación gráfica del proyecto hasta llegar al *dataset* final

Los **anunciantes**, que pueden ser tanto propietarios particulares como inmobiliarias, utilizan los portales de compraventa en el mercado inmobiliario para difundir sus **anuncios** de inmuebles. Por la creación de un anuncio, los **portales cobran** una tasa y la encarecen por otros servicios adicionales, como ocultar la dirección concreta en el caso de Fotocasa o promocionar o destacar el anuncio.

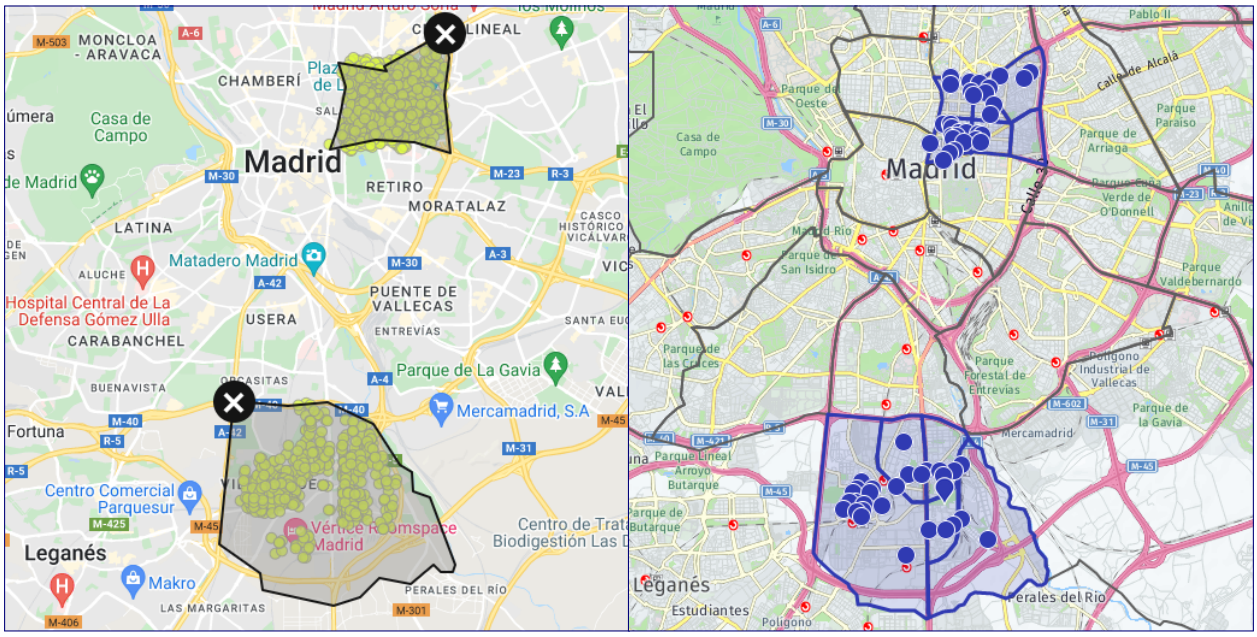
Los portales inmobiliarios ponen a disposición de los **compradores** dichos anuncios y ayudan a ponerles en contacto con los anunciantes, con los que cerrarán el trato fuera del portal.

En **este proyecto** se analizan dichos portales inmobiliarios y se vuelca en el *dataset* final la información de los anuncios sobre **viviendas en venta** en F**otocasa e Idealista**, en concreto de los distritos de **Villaverde** y **Salamanca** de Madrid, como ya se ha explicado en apartados anteriores. El resultado es un **CSV** con un registro por cada una de dichas viviendas y sus **planos** incluidos en los anuncios de Idealista, mientras que si el anuncio es de Fotocasa se han descargado todas las **fotografías** al carecer de la opción específica del plano.

5. Contenido

Los datos han sido obtenidos durante la primera semana de abril de 2022 (del 4 al 7 de abril) de los portales inmobiliarios Idealista y Fotocasa. Los registros obtenidos representan la mayor parte de las viviendas en venta en los distritos Villaverde y Salamanca en la ciudad de Madrid.

Figura 2: Distritos de Villaverde y Salamanca *scrapeados*



(a) Idealista **(b)** Fotocasa

Para obtener los datos se han utilizado las siguientes tecnologías:

- Python 3.9.10
- Selenium 4.1.3
- BeautifulSoup 4.10.0

Los datos se han obtenido mediante *web scraping*. Se ha empleado principalmente **Selenium** por ser la tecnología que mejor imita el comportamiento humano al navegar por la web, aunque también se ha utilizado **BeautifulSoup** para comprobar los códigos de error cuando la página no cargaba según lo previsto.

Para optimizar los tiempos de ejecución, el proyecto se beneficia de las **características multihilo** que ofrece Python: por un lado, cada *scraper* se lanza en su propio hilo —es decir, el *scraper* de Fotocasa y el de Idealista se ejecutarán en paralelo —y a su vez, cada *scraper* utiliza varios hilos para paralelizar el trabajo.

Por otro lado, se han utilizado esperas activas, resolución automática de *captchas* y técnicas de camuflaje mediante *proxies* y cambios de *user agent* del navegador de Selenium.

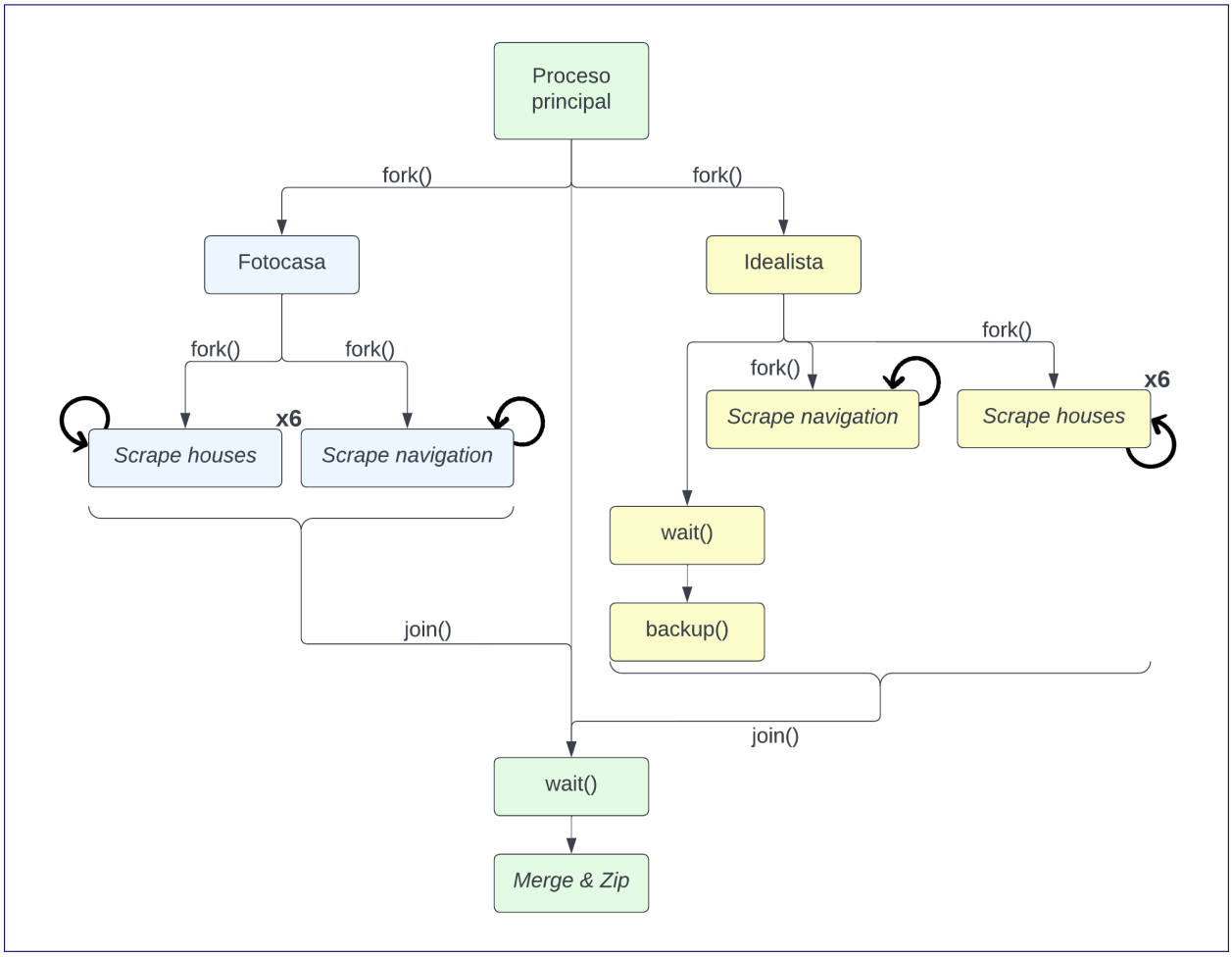


Figura 3: Hilos de ejecución del programa y *scrapers*

Respecto a las **esperas activas**, esta ha sido la técnica que mejor ha funcionado para evitar ser bloqueados por Fotocasa e Idealista, pero como punto negativo añade retardos considerables al tiempo de ejecución de los *scrapers* y no es efectiva al 100 % en Idealista.

Dada la limitada efectividad de las esperas activas en Idealista, se ha utilizado una **extensión gratuita para Google Chrome de resolución automática de *captchas*** (*Buster: Captcha Solver for Humans*), tampoco 100 % efectiva. Una alternativa consistente sería utilizar una extensión de pago de resolución de *captchas*, pero no se ha considerado por estar fuera del presupuesto del proyecto.

También se han llevado a cabo acciones para camuflar las peticiones tan repetitivas del *scraper*, con efectividad muy limitada. Por un lado, los **cambios de *user agent*** no produjeron ningún efecto en Idealista para evitar la aparición de *captchas* dado que este portal utiliza la IP del navegador para el conteo de peticiones. Sin embargo, la definición de un *user agent* válido y la **desactivación de la opción *headless*** en el navegador de Selenium evitó ser bloqueados en la primera petición a los portales.

Por otro lado, se probó la **utilización de *proxies*** para evitar los bloqueos en Idealista, utilizando *proxies* gratuitos disponibles en internet. Esta técnica, sin embargo, no terminó de ocultar la IP de la petición, por lo que no surtió ningún efecto. Sin embargo, es posible que con *proxies* de mayor calidad y, por tanto, de pago, se hubiera conseguido un resultado mucho más favorable. Finalmente, se optó por no utilizar *proxies* debido al retardo añadido a las peticiones del servidor, aunque se ha dejado esta posibilidad en el proyecto por si se quisiera retomar como opción en el futuro.

Detalles del *scraper* de Fotocasa

El primer reto solventado en el portal inmobiliario **Fotocasa** es el de la aceptación de las ***cookies***, sin lo cual no se puede navegar libremente por el portal. Por ejemplo, no permite cambiar de página y, por tanto, extraer la URL de todos los inmuebles en venta en una zona. Para saltar esta barrera, y gracias a Selenium, tras cargar la página con las viviendas por ubicación se simula el clic sobre la opción de aceptar dichas *cookies*. Por otro lado, como al abrir cada URL de cada inmueble no afecta en nada no aceptar estas condiciones, en este otro caso no se hace para no añadir retardos al proceso de *scraping*.

Además, Fotocasa cuenta con un sistema de **carga diferida** o ***lazy loading*** tanto en las páginas que ofrecen los listados de los inmuebles como en las URL de las casas en sí mismas. Esto se traduce en que los contenedores o *div* con la información de cada página se van cargando a medida que aparecen en la sección visible del navegador, es decir, conforme el usuario va navegando hacia abajo por la página (*scroll*).

Así, esta carga diferida supuso un reto a la hora de obtener la información general de cada inmueble mediante el *scraping* de las páginas de navegación y de las fotografías en las URL de las casas, ya que la información no se puede descargar tras la carga. El desafío, en este sentido, fue identificar el problema ya que el portal lo maneja de forma transparente el usuario y solo se encontró mediante depuración en un navegador aparte. Se soluciona empleando Selenium para emular el comportamiento humano de hacer *scroll* hacia abajo, de manera que se puedan cargar de forma asíncrona todos los elementos de dicha página.

Otro obstáculo solventado es que **las clases** y, en ocasiones, los ***div*** en los que se aloja la información de cada inmueble en las páginas que muestran los listados de resultados según la ubicación no son los mismos. Esto quiere decir que la estructura del portal cambia a medida que se avanza por estos listados a través de las diferentes páginas de una misma ubicación, progresión que también se hace simulando un clic sobre el botón de *página siguiente* emulando el comportamiento humano con Selenium.

Para saltar la traba de la estructura diferente, que provocaba que solo se pudiera almacenar información del 8 % de los inmuebles, se opta por evaluar la ubicación a *scrapear* en

cada caso, de manera que el programa ya detecta cuál es la estructura prevista debido a que se ha adaptado el código a cada una de ellas.

Detalles del *scraper* de Idealista

El portal inmobiliario **Idealista** cuenta con un sistema de bloqueos para *bots* utilizando *captchas* para evitar sus peticiones. Este fue el principal escollo en el diseño del *scraper*, y como se explica en esta misma sección, se solucionó parcialmente mediante cambios de *user agent*, la utilización de un *captcha solver* gratuito y esperas activas (no se utilizaron *proxies* finalmente dado que apenas producían beneficio pero introducían grandes retardos en el sistema de *scraping*).

Debido a todas las salvaguardas para evitar la detección de Idealista, el *scraper* de este portal inmobiliario dedica mucho tiempo a las esperas activas, dilatando en gran medida el tiempo requerido para obtener toda la información del sitio web.

Una consecuencia de que la ejecución completa del programa se extienda tanto en el tiempo es que la fase de pruebas es muy costosa. Así mismo, si sucede un error inesperado (dentro o fuera del programa) y el *scraper* se bloquea, con el coste que esto supone.

Teniendo estas consideraciones en mente, se ha decidido implementar un sistema de recuperación (*backup*) y reanudación del *scraping* mediante una sincronización cada `X` *links* visitados.

El **proceso de sincronización** utiliza un sistema de semáforos y eventos para controlar el número de páginas visitadas por el *scraper* de navegación los *scrapers* de inmuebles; para poder obtener una nueva página que visitar, los procesos tienen que pedir turno primero al semáforo, que solo tiene `X` huecos disponibles. Una vez los huecos del semáforo se acaban, los procesos avisan al evento y se ponen a la cola para poder obtener un nuevo *link* cuando haya hueco disponible.

El proceso principal espera a que se lance el evento para comenzar su procedimiento de sincronización: primero espera a que los procesos en curso queden a la espera de nuevos turnos en el semáforo, luego lanza el sistema de recuperación y reanudación, y finalmente libera los huecos del semáforo y da el evento como proesado y terminado.

Este **sistema de recuperación y reanudación** funciona de tal manera que se guardan en disco y se recuperan unos objetos de estado, concretamente una lista de los *links* de navegación que quedan por visitar, una lista de los *links* de inmuebles por visitar y una lista de la información de los inmuebles que hay que guardar en disco en la próxima sincronización.

Para evitar que la memoria RAM se llene antes de que el *scraping* se haya completado

—este es uno de los errores que pueden suceder, que la máquina se quede sin memoria para cargar más páginas —, solo se guarda la información de los inmuebles que no se encuentra ya salvada en disco en el fichero CSV. Cuando se realiza la sincronización, además de guardar un *pickle* de los objetos de estado se actualiza también el fichero CSV volcando la nueva información y liberando la memoria RAM asociada.

Descripción de los campos del *dataset*

A continuación se procede a describir los distintos campos del conjunto de datos final obtenido del *scraping* de Fotocasa e Idealista. Se trata de un *dataset* de 16 campos con 3.545 registros, que ocupa en total 9.8MB en disco. Cada registro puede identificarse de forma única mediante el par `(id, source)`.

- **id**: entero con el identificador numérico del inmueble, dado por Idealista o Fotocasa.
- ***source***: cadena de texto que identifica la fuente de la que se ha obtenido la información del inmueble (`"fotocasa"`/`"idealista"`).
- **url**: cadena de texto con la dirección de la página web de la que se ha obtenido la información del inmueble.
- ***title***: cadena de texto con el título del anuncio del inmueble.
- ***location***: cadena de texto con el barrio al que pertenece el inmueble (`"barrio-de-salamanca"`/`"villaverde"`).
- ***price***: entero o cadena de texto con el precio del inmueble fijado en el anuncio.
- **m2**: entero con los m^2 (metros cuadrados) del inmueble.
- ***rooms***: entero con el número de habitaciones del inmueble. El campo puede estar vacío.
- ***floor***: cadena de texto con el piso o planta del inmueble. Puede contener información adicional, como por ejemplo si se trata de un inmueble con ascensor o una orientación interior o exterior.
- ***num-photos***: entero con el número de fotografías disponibles en el anuncio.
- ***floor-plan***: entero que representa si el anuncio cuenta con un plano del inmueble (`1`) o no (`0`).
- ***view3d***: entero que representa si el anuncio cuenta con una vista en 3D del inmueble (`1`) o no (`0`).
- ***video***: entero que representa si el anuncio cuenta con un vídeo que muestra el inmueble (`1`) o no (`0`).
- ***home-staging***: entero que representa si el anuncio cuenta con la característica de *home staging* (`1`) —muestra cómo quedarían algunas habitaciones del inmueble con otro estilo, cambiando muebles, etcétera —o no (`0`).
- ***description***: cadena de texto con la descripción del anunciante para el inmueble.
- ***photo_urls***: lista con las direcciones de las fotografías del inmueble anunciado.

6. Agradecimientos

A continuación se aborda 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, este es 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 se centra 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 no se obtiene ningún tipo de información del dominio mediante *whois*.

En este sentido, la empresa es menos transparente en cuanto a cómo está **desarrollado** el portal, aunque sí que se ha 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 proyecto descrito en este informe 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 remotamente 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.

Respecto a **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 Coruña 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 se ha 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 `/es/comprar/viviendas/madrid-capital/villaverde/l?sortType=scoring`. 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 *Web scraping*[16], 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, no obstante, no se debe 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 apartados anteriores, se agradece tanto a **Idealista** como a **Fotocasa** el haber servido como soporte para el proceso de aprendizaje que se ha desarrollado durante este proyecto de captura de datos, los cuales serán empleados con un fin exclusivamente académico.

Además, se hace 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.

Por último, se agradece también a Armin Sebastian, creador de la extensión Buster: Captcha Solver for Humans, que ha hecho factible realizar *web scrapping* sobre Idealista; y al sitio web hidemy.name por su base de datos de *proxies*.

7. Inspiración

Conocer la **situación del mercado inmobiliario** es siempre interesante debido a que es un sector muy importante en España. Por ello, en muchos momentos ha servido de **termómetro** sobre la situación económica e incluso ha llegado a desencadenar la mayor crisis que se ha vivido en lo que llevamos de siglo, tanto en este país como en el mundo.

En este sentido, no es necesario extenderse mucho más puesto que en la sección 1 se puede encontrar información detallada sobre el momento en el que se desarrolla este proyecto, motivo por el que ahora resulta especialmente relevante.

El conjunto de datos resultante de este proyecto da un paso en esta dirección porque la captura se ha realizado de los dos portales inmobiliarios más empleados en España, por

lo que los datos se pueden considerar suficientemente **representativos** para los análisis posteriores.

Así, en ningún proyecto publicado hasta la fecha (o por lo menos al que haya llegado en la fase de investigación) se integra información de Idealista y de Fotocasa; tampoco existen informes que aborden las diferencias de las viviendas en venta en los **dos barrios o distritos** con los precios más extremos dentro de una localidad, lo cual es interesante justo en un momento en el que la crisis provocada por la pandemia de coronavirus ha potenciado la desigualdad, tal y como concluyen los informes de 2022 de FOESSA, Oxfam International y Observatorio Social de La Caixa.

La ciudad de **Madrid** ha sido la elegida para la captura de datos porque se trata de la más poblada de España, 3.305.408 personas a 1 de enero de 2022 según los datos del INE. Sin embargo, no existe consenso sobre si es la más cara en cuanto a la media de precio por metro cuadrado de vivienda en venta, situándose por detrás de San Sebastián y Barcelona dependiendo de la fuente consultada.

Por otro lado, a pesar de no haber fuentes oficiales actualizadas, Fotocasa e Idealista coinciden en sus informes en que el distrito más caro de la capital es el de **Salamanca** y el más barato el de **Villaverde**, por lo que estos son los que serán objeto del análisis final tras la captura realizada en esta fase del proyecto.

Una vez seleccionados los distritos, se verifica que no es útil el Anuario Estadístico de la Comunidad de Madrid (1985-2022) ni otros estudios oficiales, desactualizados o que no abordan este tipo de estudios, por lo que permitirá aportar conocimiento sobre este campo de estudio.

Teniendo en cuenta esta información, las preguntas que se pretenden resolver en este proyecto van en dos direcciones:

1. **Comparativa entre distritos**:
   - Obviando su geolocalización, ¿Cuáles son las diferencias entre las viviendas de ambos distritos? Entre las características a estudiar se encontrarían los m^2, el número de baños, el número de habitaciones, el tipo de vivienda, el año de construcción o los precios.
   - ¿Son los anuncios en un distrito por norma general más completos que en el otro? La respuesta a esta pregunta puede arrojar luz sobre si en un distrito es más sencillo vender un inmueble y, por tanto, no es necesario crear anuncios con todo tipo de detalle.
   - ¿Qué diferencias hay entre las características de cada distrito? Un análisis textual de las descripciones de los anuncios de los inmuebles de cada distrito puede ofrecer

una fotografía del distrito y mostrar las características más notables del mismo.

2. **Comparativa dentro de cada distrito**:
   - ¿Qué características comparten los inmuebles de cada distrito? Entre las propiedades a analizar se hallarían los m^2, el número de baños, el número de habitaciones, el tipo de vivienda o el año de construcción.
   - ¿Qué características son más heterogéneas entre los inmuebles de cada distrito? Las propiedades a analizar serían las mismas que en la pregunta anterior, ya que se tratan de cuestiones complementarias.
   - ¿Qué características hacen resaltar un anuncio entre los anuncios del distrito? Es decir, qué propiedad del inmueble es más cotizada o preciada en el distrito. Un análisis textual permitiría obtener más información sobre características ocultas.

Por último, cabe destacar que gracias a la capacidad de generalización del proyecto desarrollado, estas preguntas se pueden responder sobre **cualquier zona de España**, comparándolas tanto de la misma localidad como de dos diferentes simplemente cambiando las URL de inicio en las llamadas al programa. Por ejemplo, se podría haber hecho sobre el distrito más caro de Madrid y el de Barcelona, o incluso sobre todos los distritos de todas las ciudades de España.

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 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 la carpeta *house-scraper* en el siguiente repositorio de Github:

https://github.com/plazarotello/web-scraping

10. Dataset

El *dataset* con los datos de las 3.545 viviendas descritas por 16 campos en formato CSV se ha publicado en Zenodo y se puede acceder a él a partir del siguiente enlace del DOI:

https://doi.org/10.5281/zenodo.6423459

Además, como se ha especificado a lo largo del proyecto, se han descargado las fotografías de los anuncios de Fotocasa y los planos de las casas de Idealista, contenido que no ha podido compartirse en público por motivos de *copyright*.

Para facilitar que se valore la gestión de este **contenido audiovisual** en el proyecto, se ha compartido el enlace privado de Google Drive con Mireia Calvo González, profesora colaboradora a cargo del aula 2 de la asignatura Tipología y ciclo de vida de los datos de la Universitat Oberta de Catalunya (UOC).

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] Creative Commons. *Sobre las licencias*. 7 de nov. de 2017. URL: https://creativecommons.org/licenses/ (visitado 29-03-2022).

[3] 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).

[4] Instituto de Estadística de la Comunidad de Madrid. *Anuario Estadístico de la Comunidad de Madrid. 1985-2022. Urbanismo, vivienda y construcción*. Ene. de 2022. URL: http://www.madrid.org/iestadis/fijas/estructu/general/anuario/ianucap10.htm (visitado 23-03-2022).

[5] Instituto Nacional de Estadística. *Cifras oficiales de población resultantes de la revisión del Padrón municipal a 1 de enero*. 1 de ene. de 2022. URL: https://www.ine.es/jaxiT3/Datos.htm?t=2881 (visitado 27-03-2022).

[6] Instituto Nacional de Estadística. *Índice de precios de consumo. Base 2021 - Avance. Marzo 2022*. Mar. de 2022. URL: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176802&menu=ultiDatos&idp=1254735976607 (visitado 02-04-2022).

[7] Fundación FOESSA. *FOESSA presenta la primera radiografía social completa de la crisis de la COVID-19 en toda España*. 18 de ene. de 2022. URL: https://www.foessa.es/blog/foessa-presenta-la-primera-radiografia-social-completa-de-la-crisis-de-la-covid-19-en-toda-espana/ (visitado 27-03-2022).

[8] El Observatorio Social de la Fundación "la Caixa". *Radiografía de medio siglo de desigualdad en España*. Ene. de 2022. URL: https://elobservatoriosocial.fundacionlacaixa.org/documents/22890/492074/T01_ID_ES_AyalaCant%C3%B3.pdf/a0746431-109f-e009-6c77-296c378f0438?t=1642072938395 (visitado 27-03-2022).

[9] 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).

[10] 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).

[11] La Moncloa. *El Gobierno aprueba la Ley por el Derecho a la Vivienda*. 1 de feb. de 2022. URL: https://www.lamoncloa.gob.es/consejodeministros/resumenes/Paginas/2022/010222-rp_cministros.aspx (visitado 28-03-2022).

[12] OXFAM. *Las desigualdades matan*. Ene. de 2022. URL: https://oxfamilibrary.openrepository.com/bitstream/handle/10546/621341/bp-inequality-kills-170122-es.pdf (visitado 27-03-2022).

[13] Pypi. *python-whois 0.7.3*. 17 de jun. de 2020. URL: https://pypi.org/project/python-whois/ (visitado 06-04-2022).

[14] Similarweb. *Top Websites Ranking for Real Estate in Spain*. 1 de mar. de 2022. URL: https://www.similarweb.com/top-websites/spain/category/business-and-consumer-services/real-estate/ (visitado 21-03-2022).

[15] 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).

[16] 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).