Asignatura: Tipología y Ciclo de Vida de los Datos

Máster en Ciencia de Datos

PRÁCTICA 1: Web scraping

Daniel Pardo Navarro (dpardon@uoc.edu)

08 de noviembre de 2021

Índice

1. Contexto

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

Los portales inmobiliarios constituyen un elemento imprescindible en la actualidad para la compra-venta de propiedades. En este sentido, los usuarios de estos portales son tanto clientes privados como empresas y agencias inmobiliarias. Los compradores pueden hacer búsquedas parametrizando diferentes criterios de acuerdo con sus necesidades por lo que obtiene una visión global del mercado. Por lo tanto, estos portales tienen la capacidad de llegar a miles de personas y es básico para las agencias inmobiliarias tener presencia y publicitar sus activos. Estos anuncios se pueden publicar de forma gratuita o pagando una cierta cantidad con los que se obtienen diferentes beneficios (https://www.fotocasa.es/es/catalogo-productos/). En general, los portales ofrecen diferente información relativa a las características de los inmuebles como el tipo de inmueble, el precio, la ubicación o las dimensiones ya que es la información que los futuros compradores desean conocer y debe ser expuesta de manera pública con el objetivo de obtener visitas en la web.

En este proyecto se implementa un web scraper para recoger todos estos datos relevantes del portal inmobiliario *Fotocasa* (https://www.fotocasa.es/) debido a su relevancia en el mercado. Este portal es el segundo más importante de España en la actualidad, por detrás de *idealista* (https://www.idealista.com/), ya que recibe cerca de casi 16 millones de visitas al mes (https://www.helpmycash.com/cat/vender-piso/portales-inmobiliarios/).

En este aspecto, se ha descartado *idealista* ya que en sus términos y condiciones (https://www.idealista.com/ayuda/articulos/terminos-y-condiciones-generales-de-idealista/) no permiten el uso de scrapers. En referencia a las condiciones de uso y responsabilidad por el uso de la Web y Apps:

- Acceder, controlar o copiar cualquier información incluida en esta Web y apps utilizando para ello cualquier tipo de robot, spider, scraper u otro medio automático o proceso manual para cualquier propósito, sin nuestro permiso expreso y por escrito.

En referencia a la propiedad intelectual e industrial:

En concreto, no está permitido revender, realizar deep-links, utilizar, copiar, monitorizar (por ejemplo, spider, scrape), mostrar, descargar, guardar o reproducir el contenido, la información, el software, los productos o los servicios disponibles en nuestro sitio web para cualquier actividad comercial o competitiva sin autorización previa y por escrito por nuestra parte.

Por su lado, en *Fotocasa* no se hace ninguna mención ni prohibición al respecto (https://www.fotocasa.es/es/aviso-legal/ln). Además, si examinamos el archivo robots.txt (https://www.fotocasa.es/robots.txt) se observa que hay unos robots tienen permiso de acceso completo:

```
User-agent: Mediapartners-Google
Allow: /
User-agent: AdsBot-Google
Allow: /
User-agent: AdsBot-Google-Mobile
Allow: /
User-agent: TwitterBot
Allow:/
```

Otros están excluidos:

```
User-agent: Adidum
User-agent: Bloglines/3.1
User-agent: DOC
User-agent: Download Ninja
User-agent: Exabot
User-agent: Fetch
User-agent: HTTrack
User-agent: Jyxobot/1
User-agent: Mail.Ru
User-agent: Microsoft.URL.Control
User-agent: NPBot
User-agent: Offline Explorer
User-agent: SiteSnagger
User-agent: Speedy
User-agent: Teleport
User-agent: TeleportPro
User-agent: UbiCrawler
User-agent: WebCopier
User-agent: WebReaper
User-agent: WebStripper
User-agent: WebZIP
User-agent: Xenu
User-agent: Zao
User-agent: Zealbot
User-agent: ZyBORG
User-agent: cityreview
User-agent: dotbot
User-agent: e-SocietyRobot
User-agent: grub-client
User-agent: larbin
User-agent: libwww
User-agent: linko
User-agent: psbot
User-agent: sitecheck.internetseer.com
User-agent: wget
Disallow: /
```

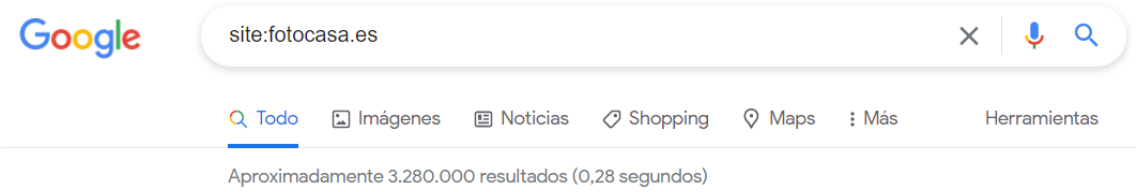
Y en mayor medida exclusión general a páginas concretas.

```
User-agent: *
Disallow: /_sysutils/
Disallow: /backwebs/
Disallow: /Backwebs/
Disallow: /buscar/
Disallow: /clientbin/
Disallow: /ClientBin/
Disallow: /complements/
Disallow: /Complements/
Disallow: /facebookconnect/
Disallow: /FacebookConnect/
Disallow: /financing/
Disallow: /Financing/
```

No obstante, todos los recursos que se utilizan para minar datos en la realización de este proyecto están permitidas. Si nos fijamos en el mapa del sitio web que se proporciona (https://www.fotocasa.es/sitemaps/sitemap_index_fotocasa_v2.xml) vemos que tiene una estructura compleja ya que para cada provincia se desarrolla uno propio.

```
▼<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  ▼<sitemap>
     <loc>https://www.fotocasa.es/sitemaps/grid/es/araba-alava_grids_sitemap_es_000000.xml.gz</loc>
     <lastmod>2021-11-04</lastmod>
   </sitemap>
  ▼<sitemap>
     <loc>https://www.fotocasa.es/sitemaps/grid/es/albacete_grids_sitemap_es_000000.xml.gz</loc>
     <lastmod>2021-11-04</lastmod>
   </sitemap>
  ▼<sitemap>
     <loc>https://www.fotocasa.es/sitemaps/grid/es/alicante_grids_sitemap_es_000000.xml.gz</loc>
     <lastmod>2021-11-04</lastmod>
   </sitemap>
```


Podemos comprobar que el tamaño de la web es grande ya que contiene más de 3 millones de enlaces.



Finalmente, podemos analizar todas las diferentes tecnologías presentes en el portal accediendo a https://builtwith.com/fotocasa.es. Por lo que respecta a información del propietario con la librería *whois* no se encuentran resultados.

2. Título

Definir un título que sea descriptivo para el dataset.

Como nombre del dataset concreto que se genera se propone: Características de las viviendas a la venta en Huesca a 4 de noviembre de 2021.

Para el nombre del archivo se elige: "comprar_viviendas_huesca_04_11_2021.csv" que también pretende ser intuitivo y dar información del contenido del dataset. En este sentido, en el código implementado el nombre de los archivos se genera de manera automática en función de las opciones indicadas y la fecha concreta en la que se ejecuta el programa. Las opciones implementadas, como veremos más adelante, son todas las que permite el portal (comprar, alquilar, compartir, etc.), en función de diferentes tipos de inmuebles (viviendas, garajes, locales, etc) y para todas las provincias españolas.

3. Descripción del dataset.

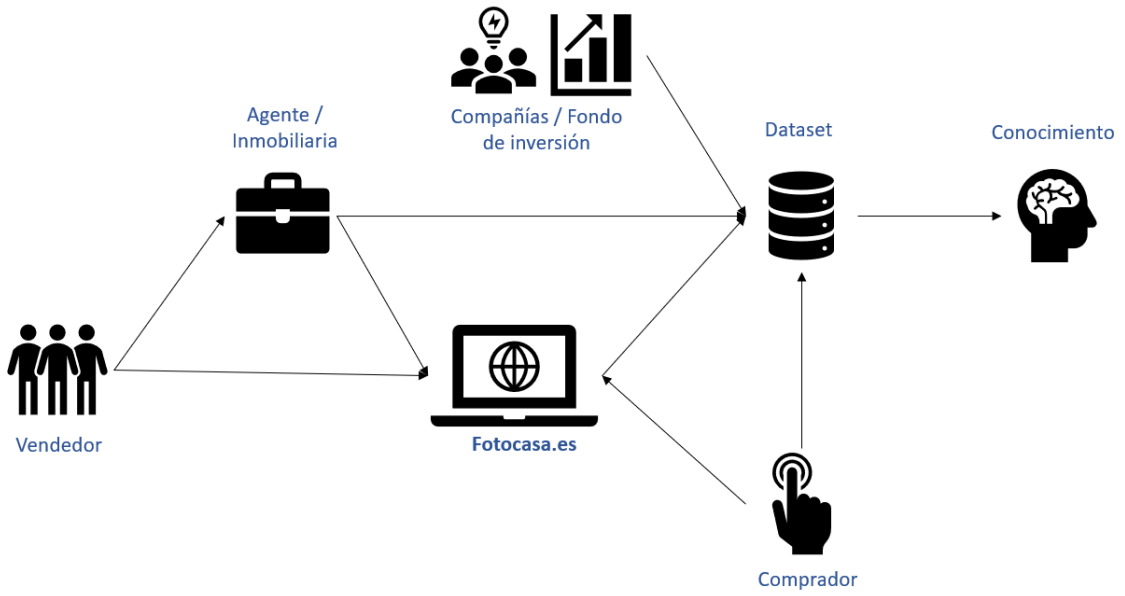
Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El conjunto de datos contiene las características principales e información relevante sobre las viviendas en venta del mercado inmobiliario de la provincia de Huesca a fecha de 4 de noviembre de 2021. Los datos se obtuvieron del portal inmobiliario *Fotocasa* (https://www.fotocasa.es/).

4. Representación gráfica

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

El siguiente diagrama muestra la utilidad del proyecto elegido. En el portal web Fotocasa.es se publican anuncios para vender o alquilar una propiedad tanto por parte de los vendedores privados como agentes o inmobiliarias. Estas publicaciones las visualizan posibles compradores en la búsqueda de alguna propiedad. Al recoger la información que contienen estos anuncios se abre la posibilidad de que el dataset generado pueda ser utilizado de múltiples formas. Por ejemplo, un comprador con suficientes conocimientos podría analizar la situación del mercado inmobiliario o hacer una búsqueda más apropiada. Sin embargo, los que pueden hacer un mayor uso de los datos son las inmobiliarias para hacer estudios de mercado o de la competencia y compañías o fondos de inversiones para buscar oportunidades de mercado. Por lo tanto, a partir del conjunto de datos y de aplicar diferentes modelos en función de los objetivos se puede generar conocimiento.



5. Contenido

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

Como ejemplo se ha utilizado la compra de viviendas en la provincia de Huesca para crear el conjunto de datos. Esta elección se ha hecho porque la compra de viviendas es la acción de búsqueda más habitual en estos portales y por razones de tiempo de ejecución debido al número de viviendas disponibles en el portal para esta provincia (algo más de 2000).

No obstante, el programa generado permite generar conjuntos de datos para todas las posibilidades que existen en el portal *Fotocasa* y que se describen a continuación:

Para la **acción** que se desea realizar podemos escoger entre compra, alquiler, alquiler vacacional, compartir o promociones de obra nueva. Cabe destacar que para las acciones de alquiler vacacional, compartir y promociones de obra nueva el tipo de inmueble solo puede ser viviendas.

La segunda opción para seleccionar es el **tipo** de inmueble. En este caso, los inmuebles pueden ser viviendas, locales, garajes, edificios, oficinas, trasteros o terrenos.

Finalmente se selecciona la **provincia** sobre la que se quieren obtener los datos. Se tienen en cuenta las 50 provincias españolas además de las ciudades autónomas de Ceuta y Melilla.

Además del archivo .csv, se obtiene una carpeta "images" donde se almacena la foto principal de cada publicación. Cada imagen tiene como nombre el número de identificador que se relaciona con el conjunto de datos.

El dataset incluye 22 campos diferentes, aunque no todos ellos son obligatorios. A continuación, se describe cada uno de los campos:

- **Identificador** → Como obligatorio. Este campo contiene un identificador único que permite identificar a las publicaciones de forma unívoca. Se utiliza el mismo identificador que asigna el portal web.
- **Provincia** → Campo obligatorio. Nombre de la provincia en la que se ubica la propiedad.
- **Municipio** → Campo obligatorio. Municipio donde se encuentra el inmueble.
- **Zona** → Campo obligatorio. Zona de la localidad del inmueble.
- **Tipo** → Campo obligatorio. Permite identificar el tipo de inmueble concreto. El tipo de inmueble puede ser en general piso, loft, apartamento, ático, dúplex, planta baja, estudio, casa o chalet, casa adosada, finca rústica, garaje, oficina, trastero, terreno, edificio, local o nave industrial dependiendo de las opciones seleccionas.
- **Precio** → Campo obligatorio. Precio de venta expresado en euros y sin comas ni puntos decimales.
- **Dimensión** → Campo obligatorio. Dimensión del inmueble expresada en metros cuadrados.
- **Habitaciones** → Número de unidades de habitaciones con las que cuenta el inmueble.
- **Lavabos** → Número de unidades de lavabos de la propiedad.
- **Planta** → Número de planta en la que se encuentra el inmueble.
- **Ascensor** → Si la propiedad cuenta con ascensor el campo será "SI". Para el resto de los casos el campo estará vacío ya que no se puede deducir si realmente no tiene ascensor o falta indicar esta información.
- **Parking** → Especifica si la propiedad cuenta con parking.
- **Estado** → Estado de conservación del inmueble.
- **Calefacción** → Tipo de energía que utiliza el mecanismo de calefacción.
- **Agua caliente** → Tipo de energía que se utiliza para calentar el agua.
- **Antigüedad** → Años de antigüedad desde la construcción el inmueble.
- **Orientación** → Orientación respecto de los puntos cardinales.
- **Amueblado** → Contiene la información sobre si el inmueble está amueblado.
- **Consumo** → Calificación energética de consumo de energía.
- **Emisiones** → Calificación sobre las emisiones de CO_2.
- **Vendedor** → Identifica si es un particular o el nombre de la agencia inmobiliaria que publica el anuncio.

- **Imagen** → Campo obligatorio. Se identifica con "SI" en el caso de que se haya descargado la imagen principal del anuncio. En caso de que el anuncio no contenga imagen el campo será "NO". Las imágenes se guardan en el fichero "images" y tiene como nombre el mismo valor del identificador de la publicación concreta.

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 actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El propietario del conjunto de datos es Fotocasa. Se define cono un portal inmobiliario líder de España fundado en 1999. La compañía pertenece a Adevinta, una compañía líder en marketplaces digitales y una de las principales empresas del sector tecnológico del país con portales como habitaclia, Infojobs.net, coches.net, motos.net o Milanuncios (https://www.fotocasa.es/es/quienes-somos/).

Por lo que respecta a análisis similares, uno de los datasets de aprendizaje más famosos es el de *Boston house-price* que data del año 1978 sobre el que se han realizado múltiples estudios y análisis para predecir el valor de una vivienda en Boston. Para ello, se han utilizado una gran variedad de técnicas desde el análisis de regresión (https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d) hasta las redes neuronales (https://www.thepythonacademy.com/post/predicting-house-prices-using-a-deep-neural-network-case-with-the-boston-dataset) e incluso existen competiciones al respecto (https://www.kaggle.com/c/house-prices-advanced-regression-techniques). Sin embargo, este conjunto de datos cuenta con variables adicionales como el índice de criminalidad o los impuestos.

Existen multitud de casos de estudio para predecir el valor de una vivienda. En general, las publicaciones más recientes utilizan técnicas de machine learning para generar los modelos con (https://www.sciencedirect.com/science/article/pii/S1877050920316318) o determinar que características son las más importantes y tienen un mayor peso de influencia en el precio de una vivienda (https://ciencia.lasalle.edu.co/eq/vol1/iss30/1/).

Sin embargo, los conjuntos de datos de este tipo también permiten realizar estudios retrospectivos que permitan estudiar el impacto de diferentes acontecimientos en relación con el precio de una vivienda (https://recyt.fecyt.es/index.php/CyTET/article/view/86498/63375).

Por lo que respecta a los pasos seguidos en cuanto a principios éticos y legales se ha respetado la voluntad expresada por los propietarios de los portales web. Tal y como se ha desarrollado en el apartado 1, se ha descartado el portal *idealista* debido a que expresaba su voluntad de no usar bots contra los datos de su web. Por este motivo se ha utilizado el portal *Fotocasa* que no prohíbe este tipo de prácticas. Además, se ha revisado el archivo robots.txt para estar seguros de que no se incumplían los deseos de los propietarios de los datos en cuenta al uso de bots. Por otro lado, con el objetivo de evitar el uso de información personal o sensible se ha evitado recoger cierto tipo de datos. En este sentido, en el portal web podemos encontrar datos de la ubicación del inmueble, una descripción o el número de teléfono de contacto. Se ha considerado

que este tipo de información no es relevante para este conjunto de datos y que, además, se trata de información personal por lo no ha sido almacenada.

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.

Este conjunto de datos, o la posibilidad de analizar un conjunto de datos de cualquier provincia en general con el código implementado, tiene múltiples utilidades que pueden ser interesantes a diferentes tipos de usuarios. El sector inmobiliario y de la vivienda es ampliamente estudiado y motivo de debate político. Por lo tanto, conocer el mercado inmobiliario de provincias o zonas concreta, según el interés, puede ayudar a tomar alguna decisión política determinada. En este sentido, algo que podría complementar y enriquecer el dataset actual es añadir diferente información relevante para cada municipio tal y como se hace en el conjunto de datos de Boston mencionado en el apartado anterior. Por ejemplo, se pueden incluir datos de desempleo del municipio o índices de criminalidad para poder entender mejor los datos del precio de la vivienda.

Otro punto interesante de este tipo de conjuntos de datos es el uso que pueden darles las agencias inmobiliarias para hacer un estudio del mercado actual y poder fijar los precios en base a la competencia. También permitiría conocer la influencia y las dimensiones de la propia empresa en función de la cantidad de propiedades que tiene una en cartera determinada inmobiliaria. Los estudios de mercado también pueden ser útiles para distintas compañías de diferentes tipos. Por ejemplo, si se detecta que en una zona de forma significativa los inmuebles son antiguos o que no cuentan con calefacción o aire acondicionado empresas de estos sectores podrían utilizar la información para promocionarse o crear ofertas. Del mismo modo, los fondos de inversión pueden utilizar estos datos para detectar oportunidades de mercado donde invertir y comprar diferentes tipos de propiedades.

En general, algunas de las preguntas que se pueden responder con este conjunto de datos son: ¿Qué municipio tiene la vivienda más cara?, ¿Qué características son importantes a la hora de determinar el precio de la vivienda?, ¿Cuál es el precio adecuado de una vivienda en una zona y con unas características determinadas?, ¿Qué vivienda está por debajo de su valor? o ¿Qué produce que una vivienda esté por encima o por debajo del valor de mercado de su zona? entre otras posibles cuestiones que puedan surgir. Además, como se ha comentado se puede enriquecer el conjunto de datos con información adicional para responder a preguntas más complejas dependiendo de las necesidades.

8. Licencia

Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección.

El dataset se encuentra publicado bajo la licencia Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/legalcode.es). Con esta licencia se

puede compartir, copiar y redistribuir el material en cualquier medio o formato. Además, se permite adaptar, remezclar, transformar y construir a partir del dataset original según las necesidades para cualquier propósito, incluso comercialmente. De esta forma, se permitiría enriquecer el conjunto de datos añadiendo nuevas variables que puedan ser interesantes como se ha mencionado en el apartado anterior.

Por otro lado, con esta licencia se requiere hacer atribución y dar crédito de manera adecuada al autor, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Esto debe hacerse de forma razonable y sin sugerir que se tiene apoyo de la licenciante.

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 se encuentra disponible en el siguiente repositorio:

https://github.com/Daniel-Pardo/webScraping

10. Dataset

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

Los enlaces para acceder al conjunto de datos son los siguientes:

https://zenodo.org/record/5645835

https://doi.org/10.5281/zenodo.5645835