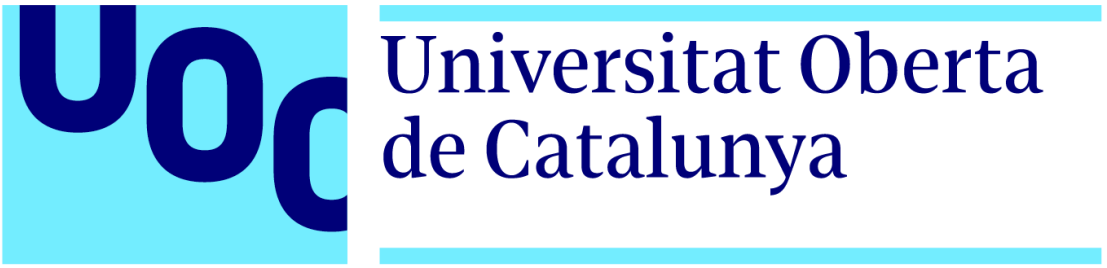
Tipología y ciclo de vida de los datos

Práctica 1 – Web scraping



Máster en Ciencia de Datos

Francisco de Borja Navas Torres

Álvaro de la Fuente Díaz

Abril 2021

Índice

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

Nos hemos puesto en la piel de una pequeña promotora que se ha incorporado recientemente en el mercado inmobiliario y que necesita apoyo e instrumentos que le permita identificar, por un lado, qué suelos son interesantes para invertir y, por otro lado, qué tipologías de viviendas son las más recomendadas de construir para maximizar los beneficios.

En primer lugar, para ayudar a esta promotora hemos decidido hacer un estudio de benchmarking con el objetivo de conocer aspectos clave de la competencia. Para ello, nos ha parecido interesante basarnos en la técnica de "web scraping". Mediante esta, se pretende obtener información de la competencia acerca de dónde invierten, qué tipo de viviendas construyen, y que despierta el interés de los clientes con respecto a estas viviendas.

La promotora seleccionada para llevar a cabo esta técnica es Metrovacesa, una de las principales promotoras inmobiliarias del mercado nacional. Realizando una inspección visual sobre el código fuente de la página web, podemos ver cómo está estructurada la información de cada promoción mediante una tabla. A partir de ella se pueden obtener características de las promociones, de las viviendas disponibles y los precios de estas. Por lo tanto, dentro del contexto en el que nos encontramos y las alternativas que hemos estado analizado, damos como válida la opción de Metrovacesa, ya que podemos obtener la información que necesitamos.

Adicionalmente, hemos verificado la estructura del fichero robots.txt, asegurándonos de que no violamos ningún principio del propietario de la web.

```
User-agent: *
Disallow: /wp-snapshots/
Disallow: /wp-content/uploads/wpcf7-submissionsOLD/
Disallow: /test/*
Disallow: /cgi-bin
Disallow: /*/attachment/
Disallow: /tag/*
Disallow: /xmlrpc.php
Disallow: /?attachment_id*
Disallow: /*.pdf$
Disallow: */cfdb7_uploads/
Disallow: */wpcf7-submissions/
Disallow: /crm/*
```

Por tanto, nos aseguramos que no entramos en los directorios que se marcan como no accesibles. No es necesario aplicar un delay entre peticiones html ya que no se especifica en este fichero.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

El título del dataset lo hemos definido como "Metrovacesa - Unidades inmobiliarias en comercialización".

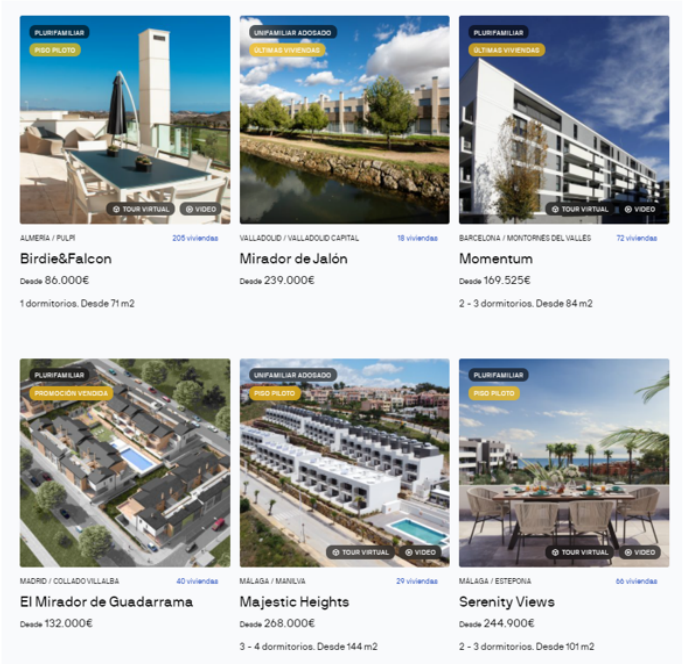
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

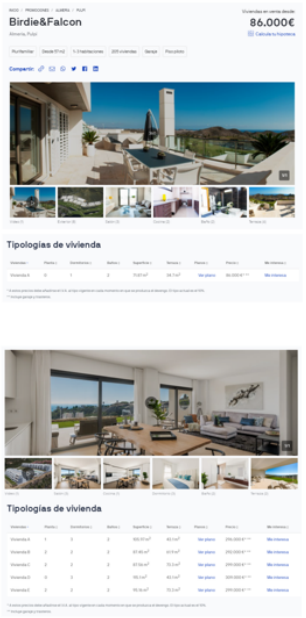
El dataset contiene información de cada una de las diferentes unidades inmobiliarias que tiene en comercialización la promotora Metrovacesa por cada una de sus promociones, municipios y provincias. Dentro de cada una de estas unidades, obtenemos los datos básicos de cada una de ellas y el precio al que está comercializado.

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

La siguiente figura representa el esquema del proyecto. En primer lugar, se encuentra el contenido de la página web principal de la promotora Metrovacesa, de donde extraemos el árbol de links por provincia que nos lleva a un segundo paso donde por cada provincia disponemos de los links de cada una de las promociones. En tercer lugar accedemos al detalle de la promoción, donde a través de un proceso de web scraping se extrae la información clave para su posterior análisis. Esta información es exportada en un fichero csv formando el dataset "Metrovacesa - Unidades inmobiliarias en comercialización".

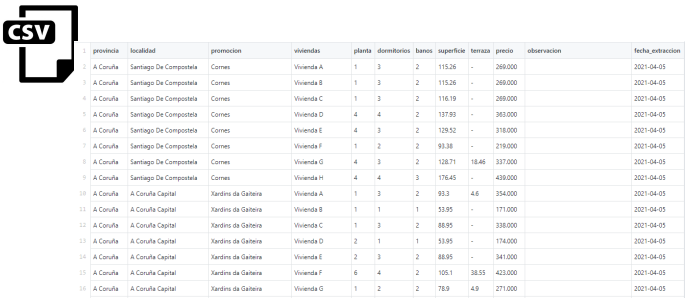
Página web "Metrovacesa". Links de promociones por provincia.





Página de promociones por provincia. Links por promoción.

Información objeto de análisis web "Metrovacesa"



Dataset "Metrovacesa - Unidades inmobiliarias en comercialización"

Proceso de web scraping

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

El dataset está compuesto de los siguientes campos:

- **provincia**: provincia en la que se encuentra la promoción.
- **localidad**: localidad en la que se encuentra la promoción.
- **promocion**: nombre comercial de la promoción.
- **viviendas**: tipo de vivienda dentro de la promoción.
- **planta**: planta en la que se encuentra la vivienda.
- **dormitorios**: número de dormitorios de la vivienda.
- **baños**: número de baños de la vivienda.
- **superficie**: superficie expresada en metros cuadrados de la vivienda.
- **terraza**: superficie expresada en metros cuadrados de la terraza de la vivienda.
- **precio**: precio de la vivienda.
- **observacion**: comentarios adicionales. Por ejemplo, hay promociones que han sido publicadas pero sin datos sobre ellas al ser muy recientes. Para estos casos en este campo se incluye “Promoción disponible próximamente”.
- **fecha_extraccion**: fecha en la que se ha realizado la extracción de la información.

Los datos almacenados en el dataset han sido extraídos de la página web de la promotora Metrovacesa: https://metrovacesa.com. En esta página web, la promotora publica las diferentes promociones que está realizando por todo el territorio nacional. De esta forma, se van actualizando los datos en la web a medida que surgen nuevas oportunidades inmobiliarias o se venden las que estaban disponibles. Por ello, el tiempo de validez de los datos está comprendido entre la fecha de lanzamiento de la promoción inmobiliaria y la fecha de venta de todos los inmuebles asociados a la promoción.

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.

Los datos almacenados en el dataset que se presenta han sido obtenidos de la página web de la promotora inmobiliaria Metrovacesa (https://metrovacesa.com). Su extracción se ha llevado a cabo utilizando técnicas de Web Scraping a través de un script codificado en lenguaje de programación Python.

La elaboración del dataset, se ha basado en la idea del contenido del conjunto de datos “House Prices - Advanced Regression Techniques” ubicado en el directorio web https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data. En este dataset se detallan diferentes características de viviendas ubicadas en Iowa, tales como el precio de venta, la clase del edificio o la superficie.

7. Inspiración. Explique 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.

Con el objetivo de identificar qué suelos son interesantes para invertir y qué tipologías de viviendas son las más recomendadas de construir para maximizar los beneficios se ha procedido a realizar un estudio de la competencia. Para ello se ha seleccionado la inmobiliaria "Metrovacesa". De esta, se han analizado sus promociones inmobiliarias, el tipo de viviendas que construyen y que despierta el interés de los clientes con respecto a estas viviendas.
Del mismo modo que el dataset "House Prices - Advanced Regression Techniques" detalla las principales características de un conjunto de viviendas de Iowa, en este caso se ha pretendido almacenar en un mismo fichero las características de los inmuebles que se comercializan en las diferentes promociones que la empresa Metrovacesa está llevando a cabo por todo el territorio nacional.

Adicionalmente, con los datos recabados de la web de Metrovacesa se pueden aplicar análisis muy interesantes para calcular las distribuciones óptimas de las promociones que vas a construir. Esto lo podríamos hacer almacenando el histórico de varios meses de la web de metrovacesa y calculando de cada promoción qué unidades son las que primero se venden y a qué precio, por lo que dispondremos de la información para adaptar la estrategia de la promotora al mercado.

8. Licencia. Seleccione una de estas licencias para su dataset y explique 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**

La publicación del dataset está regida por la licencia **CC0: Public Domain License**. Bajo esta licencia, se da al conjunto de datos de un carácter de dominio público renunciando a todos los derechos de la obra por parte de los creadores.

El dataset se puede copiar, modificar, distribuir e interpretar para fines comerciales si se deseara sin necesidad de pedir permiso para ello.

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

https://github.com/fnavast/MetrovacesaScrapingWeb

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El dataset está almacenado en el repositorio Zenodo en la siguiente dirección: https://zenodo.org/record/4678321#.YHGfCugzaUk

Al publicarlo en dicha plataforma, se le ha asignado de manera automática el DOI "10.5281/zenodo.4678321".

11. Contribuciones al trabajo

En la siguiente tabla se detallan los nombres de las personas que han participado en el desarrollo de las tareas de análisis y extracción de los datos, así como en la elaboración del fichero pdf en el que se detalla el contenido del proyecto.

Contribuciones	**Firma**
Investigación previa	FBNT y AFD
Redacción de las respuestas	FBNT y AFD
Desarrollo del código	FBNT y AFD