Tipología y Ciclo de vida de los datos

# PRA1: Web Scrapping

### Autores

Rafael Jimenez Sarmentero

Jorge Marchán Gutiérrez

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

**En nuestra primera reunión vía Google Meet, lo primero que nos cuestionamos fue sobre qué área o materia iba a versar nuestra práctica. La intención era sugerir algunas propuestas y optar finalmente por una de ellas.**

**Como uno de los miembros del equipo trabaja actualmente en el sector inmobiliario, la idea de enfocarnos en este ámbito surgió enseguida.**

**El primer portal que nos vino a la mente, como no podía ser de otra manera, fue Idealista.**

**Escrutamos su fichero robots.txt y no observamos que estuviera desaconsejada la extracción automática de datos. Sin embargo, una lectura de las «Condiciones generales» nos disuadió de tal intención.**

**De modo que nuestra reunión continuó con una búsqueda de portales inmobiliarios que nos permitieran hacer el *scraping* de manera legal.**

**Así, tras algunas búsquedas infructuosas, llegamos a Fotocasa <**<https://www.fotocasa.es/es/>**>.**

**Una inspección detallada nos hizo reparar en que, a través de la página principal podíamos extraer los enlaces a las diferentes viviendas. Asimismo, observamos que cada vivienda ofrecía una página de detalle en la que podían obtenerse distintos campos que serían de interés para la construcción de un *dataset:* área de la vivienda, número de habitaciones, número de cuartos de baño, etc.**

**2. Título. Definir un título que sea descriptivo para el dataset.**

**«madrid\_real\_state\_pricing.»**

**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 corresponde con los datos de un conjunto de viviendas de Madrid, concretamente las tres primeras páginas de la búsqueda ya que según su robots.txt, Fotocasa solo permite búsquedas sobre las tres primeras páginas que conforman una búsqueda, sin embargo, el script se ha preparado para que se le pueda pasar el número de páginas que deseas obtener y el término de búsqueda, por lo tanto, podemos buscar para Madrid o para cualquier provincia o ciudad.**

**El dataset contiene información relativa al precio de la vivienda y sus características más básicas y que según nuestro criterio más pueden influir en este precio.**

**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 como se han recogido.**

**Los campos que conforman el *dataset* extraído son los siguientes:**

* **price: Precio de la vivienda.**
* **bedrooms: Número de dormitorios.**
* **bathrooms: Número de cuartos de baño.**
* **sqm: Metros cuadrados de la vivienda.**
* **floor: Número de planta en la que se halla la vivienda.**
* **type: Tipo de la vivienda (piso, chalet, etc.).**
* **orientation: Orientación de la vivienda (norte, sur, este, oeste).**
* **condition: Estado de la vivienda (a reformar, etc.).**
* **hot water: Forma de suministro del agua caliente.**
* **elevator: Si la vivienda dispone o no de ascensor.**

**En lo referente al periodo de tiempo de los datos es marzo de 2022, es decir, en el momento que se lanza el script el periodo de los datos con el cual se genera el dataset es “momento actual”.**

**En cuanto a la forma de recolección es navegando por los distintos resultados de las tres primeras páginas de búsqueda y accediendo a las distintas fichas de alojamiento para a continuación extraer los campos que hemos mencionado anteriormente, todo esto de manera automática, indicándole término de búsqueda y número de páginas**