PRAC 1: WEB SCRAPING

Tipología y Ciclo de Vida de los Datos

Máster en ciencia de datos

Jaime Gimeno Ferrer y Reynel López Lantigua

PREGUNTAS

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

Este trabajo se hace como parte de la asignatura de Tipología y Ciclo de Vida de los Datos del Máster en Ciencia de Datos de la Universidad Oberta de Catalunya. El objetivo es encontrar una página web para practicar *web scraping* sobre esta y afianzar así los conceptos estudiados.

Más que buscar un tipo de información en concreto, nos centramos en la búsqueda de un sitio web sobre el que fuésemos capaces de aplicar el *web scraping* y que de la misma manera **nos supusiera un reto** con el cual aprendiéramos a fondo.

Hemos elegido la web de **coches.com** debido a que, tras meditar y probar con muchas páginas, hemos encontrado en esta una **gran oportunidad para aprender de forma más avanzada *web scraping* teniendo que enfrentarnos a diversos obstáculos**. Entre estos obstáculos vimos una "trampa de araña" que generaba infinitas páginas web de coches por cada marca y un persistente *pop up* para aceptar las cookies que nos dio bastantes problemas.

Además de encontrar en este sitio web las características que buscábamos, los datos que encontrábamos en el mismo tenían una **estructura perfecta para organizarla en un *csv*** en forma de *dataset*.

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

El título que hemos elegido para el *dataset* es:

COCHES A LA VENTA EN COCHES.COM HASTA 2020

Para el nombre del dataset como archivo hemos escogido *cars_for_sale_coches_com.csv.*

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* que hemos conseguido obtener **recopila los datos de todos los coches de KM0 y de segunda manos** ofertados en la página web www.coches.com con sus diferentes atributos, desde el año del coche, pasando por el precio, hasta los caballos. Vemos los diferentes atributos en el siguiente punto.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

El dataset generado tiene una serie de atributos que podríamos representarlos como en la Figura 1 si la tabla fuese parte de una base de datos relacional (esquema entidad-relación con solo atributos en este caso).

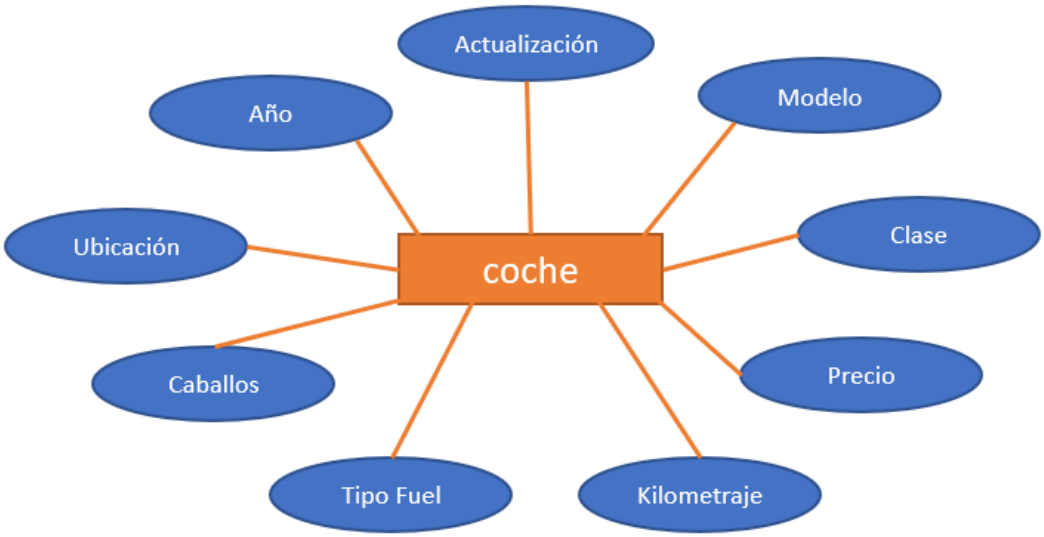

Fig. 1 Atributos de la entidad coche

Una buena forma de visualizar el dataset podemos verla en la Figura 2, donde se muestran los primeros datos de este.

Índice	last_update	model	class	price	km	fuel	cv	location	year
0	last_update	model	class	price	km	fuel	cv	location	year
1	5 días	Aston Martin 5.2 608	Km 0	165.900	6.000	Gasolina	608	Madrid	2019
2	11 meses	Aston Martin 4.0 510	Km 0	240.000	7.000	Gasolina	510	Madrid	2019
3	13 meses	Aston Martin 5.2 608	Km 0	173.297	100	Gasolina	608	Madrid	2019
4	4 horas	Audi Q5 55 TFSIe S line	Km 0	49.500	4.500	Híbrido	367	Granada	2020
5	18 horas	Audi A6 Allroad 45 TDI	Km 0	62.900	8.000	Diesel	231	Guipuzcoa	2019
6	5 horas	Audi Q3 35 TFSI Advanced	Km 0	33.500	20	Gasolina	150	Granada	2020
7	10 horas	Audi A1 Sportback 30 TFSI	Km 0	22.800	9.129	Gasolina	116	Toledo	2019

Fig. 2 Primeras muestras del dataset

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

Los datos que incluimos en el dataset son los que aparecen en la vista previa de cada coche si vamos navegando por las distintas marcas (por cada marca tenemos muchas páginas igualmente). Hemos escogido estos datos ya que los consideramos suficientemente ilustrativos. Dentro de cada página de cada coche aparece algún dato más que no consideramos muy relevante y hemos obviado su recolección debido a que hemos visto inviable el recorrer las 110.000 páginas web con el tiempo que disponíamos (una por coche). Recorriendo las páginas y centrándonos en los datos de las vistas previas nos ahorramos mucho tiempo obteniendo resultados muy similares.

A continuación, enumeramos los campos que incluye el dataset.

- **last_update**: Actualización de los datos
- **model**: Modelo del coche, marca y nombre

- **class**: Clase del coche, puede tomar valor “Km 0” o “Segunda mano”
- **price**: Precio del coche
- **km**: Kilómetros recorridos por el coche
- **fuel**: Tipo de fuel que utiliza
- **cv**: Caballos medidos en cv
- **location**: Localización del coche
- **year**: Año de lanzamiento del coche

En cuanto al periodo de los datos tenemos que decir que estos datos son válidos para la fecha de hoy 5 de octubre de 2020. Esto es por dos motivos, el primero y más claro es que hemos **recolectado los datos hasta hoy** y habrá que ir actualizándolos conforme se suban nuevos coches. El segundo motivo es que **los coches pueden irse vendiendo**, en cuyo caso desaparecerán de la vista.

Esto no significa que los datos no tengan valor, en la mayoría de casos de uso de estos datos, desde estudios estadísticos hasta modelados con ML, los datos no necesitan estar actualizados al 100% y si fuese así, nos sirve para hacer estudios de los datos hasta la fecha.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigacion o analisis anteriores (si los hay).

Agradecemos a los creadores de la página web en cuestión el hecho de que dejen la posibilidad de hacer *web scraping* sobre sus datos (tal y como podemos ver en su archivo robots.txt). Del mismo modo, este proyecto habría sido muy distinto sin el uso de la librería Selenium creada por Jason Huggins la cual nos ha facilitado mucho el navegar y recopilar datos de la web.

7. inspiración. Explique por qué es interesante este conjunto de datos y que preguntas se pretenden responder.

Encontramos interesante este dataset debido a distintas razones. La primera razón es que es un *dataset* que como bien sabemos ha sido **generado a partir *web scraping* y que hasta ahora no existía**, o al menos no hemos podido encontrarlo en internet.

Además, se nos han ocurrido **muchos estudios** que podrían llevarse a cabo con el uso de este dataset. Entre ellos están: El estudio de cómo los distintos atributos influyen al precio, kilometraje, caballos… De esta forma podríamos generar un **modelo que pueda predecir el precio** de un coche que quieras vender basándose en sus características. Otro estudio podría ser cómo han ido **evolucionando los coches**, viendo la predominancia del tipo de combustible, qué modelos han sido los más vendidos, etc. Muchos otros estudios pueden realizarse a partir de estos datos.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección.

Hemos escogido esta licencia para nuestro dataset ya que posee las siguientes características que podemos ver en https://creativecommons.org/licenses/by-nc-sa/4.0/ :

- Permite **compartir**, copiar y redistribuir el dataset en cualquier medio y formato.
- **Adaptar**, transformar y construir sobre el mismo.

Todo esto mientras se de **crédito** a los autores, adjuntar un link a la licencia e indicar si se han hecho cambios sobre el *dataset.* Si se modifica, ha de usarse la misma licencia.

Además, **no permite usarse para objetivos comerciales**, solo de estudio.

Estas condiciones nos parecen las más apropiadas para nuestro *dataset*.

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

10. Dataset. Publicacion del dataset en formato CSV en Zenodo (obtencion del DOI) con una breve descripcion.