Tipologia i cicle de vida de les dades

# PRACTICA 1 – WEB SCRAPING

Julia Soler ([jsolerni@uoc.edu](mailto:jsolerni@uoc.edu)) & Antonio Castro ([acastrom@uoc.edu](mailto:acastrom@uoc.edu))

## Punto 1

### Contexto

Uno de los integrantes del equipo, Toni Castro, está interesado en vender su actual coche.

Se pretende realizar web scraping a una página de compra/venta de coches,con el objetivo de captar los datos de vehículos similares y poder extraer información útil para su venta.

Por este motivo, el fin de presente práctica, es generar un fichero, que contiene los datos de coches anunciados en una web de segundo mano, para aportar información a un posible vendedor,siendo el input principal , el modelo y marca del coche que se desea vender.

Existen muchas Webs de compra/venta de vehículos; se ha elegido la página <https://www.autoscout24.es/>, porque permite Web scraping.

En su fichero robots.txt (anexado) se indica :

Allow: /home/index/offer.asp

Cómo comentario, Google nos dice que este site tiene mas de 6.6 millones de links.

## Punto 2

### Título Dataset

El titulo del fichero de datos es Datos\_coches\_ventas.csv

## Punto 3

### Descripcion del dataset

El fichero presenta un listado de atributos de coches anunciados para su venta, a partir de parámetros de entrada: marca/modelo/año/versión.

## Punto 4

### Esquema

El esquema se realizará en la entrega final de la práctica

## Punto 5

### Contenido

La idea es extraer todos los datos relevantes, que puedan influir en la venta de un coche.

Se generará un fichero .csv con los siguientes campos:

* **Fecha extracción datos**
* **Precio**: PVP de vehículo al contado
* **Año**: Año/mes de la matriculación
* **Marca**: Fabricante del vehículo
* **modelo**: Modelo
* **Versión**: Versión detallada del vehículo
* **Combustible**: Gasolina, diésel, eléctrico
* **Kilómetros**: Kilómetros recorridos
* **Ciudad**: Lugar donde se encuentra el vehículo

Periocidad de extracción: La periocidad de la extracción será semanal.

Método de extracción: Script de Python y almacenamiento local en ficheros csv.

## Punto 6

### Agradecimientos

La consulta 'whois' a esta web, no nos proporciona información.

>>> print(whois.whois('https://www.autoscout24.es')) { "emails": null, "state": null, "status": null, "country": null, "domain\_name": null, "creation\_date": null, "registrar": null, "dnssec": null, "updated\_date": null, "name\_servers": null, "expiration\_date": null, "whois\_server": null, "referral\_url": null, "org": null, "address": null, "city": null, "zipcode": null, "name": null }

## Punto 7

### Inspiración

Nos permite obtener información sobre vehículos semenjantes al que se quiere vender y en qué rango de precios se mueven.

Con los datos del fichero se podria extraer información útil para un venta.

Por ejemplo:

* Cantidad de coches que se venden de ese modelo.
* Rango de preciós.
* Antigüedades medias del mismo modelo.
* Kilometraje medio.

## Punto 8

### Licencia

<https://es.wikipedia.org/wiki/Licencias_Creative_Commons>

Yo optaría por esta, por ser la permite un mayor uso no comercial.

|  |  |  |
| --- | --- | --- |
| Atribución-No Comercial-Compartir Igual | (CC BY-NC-SA) |  |

## Punto 9

Copiar código fuente

## Punto 10

Se sube nuestro dataset al repositorio ZENODO (sincronizado con Github) y se obtiene su DOI

## Punto 11

Se presenta la practica 1 en GitHub

<https://github.com/afcastrom/TCVD-PRA1>

## Contribuciones

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Búsqueda previa | Toni Castro y Julia Soler |
| Redacción de las respuestas | Toni Castro y Julia Soler |
| Desarrollo del código | Toni Castro y Julia Soler |

## Bibliografía