

Tipologia i cicle de vida de les dades

PRACTICA 1 – WEB SCRAPING

Julia Soler (jsolerni@uoc.edu) & Antonio Castro (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 de anunciantes particulares 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

Como comentario, Google nos dice que este site tiene más de 6.6 millones de links.

Punto 2

Título Dataset

El título del fichero de datos es Datos_coches_ventas.csv

Punto 3

Descripción del dataset

El fichero presenta un listado de datos de vehículos de particulares, anunciados para su venta, a partir de los siguientes parámetros de entrada: marca/modelo/año/versión...

Punto 4

Esquema

Punto 5

Contenido

Se pretende 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

Periodicidad de extracción: La periodicidad de la extracción será semanal. Se considera que la información no cambia mucho diariamente.

Método de extracción: Se ha realizado mediante Script de Python utilizando la librería de Python para BeautifulSoup y almacenamiento local en ficheros .csv.

Punto 6

Agradecimientos

La consulta 'whois' a esta web utilizada para realizar scraping, www.autoscout24.es 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 }
```

Resultado de búsquedas de proyectos parecidos:

<https://statisquo.de/2020/01/16/autoscout24-mining-webscraping-mit-python/>

<https://github.com/mauropelucchi/autoscout24>

Bibliografía

1. *Página utilizada para extraer la información de los coches de particulares a la venta*
<https://www.autoscout24.es/>
2. Moya R. (2015) *Scraping en Python (BeautifulSoup), con ejemplos*.
<https://jarroba.com/scraping-python-beautifulsoup-ejemplos/>
3. Subirats L. Calvo M. *Web Scraping*. Recurso proporcionado en el temario de la asignatura.

Punto 7

Inspiración

Para un particular que quiere vender su coche, es interesante obtener de forma periódica y ágil, los datos de coches particulares ya ofertados para su venta, del mismo modelo y versión que se desea vender.


Con el script implementando, el usuario podrá obtener valiosa información sobre coches en venta de particulares, semejantes al que se quiere vender.

- Cantidad de coches que se venden de ese modelo/Versión.
- Rango de precios.
- Antigüedades medias del mismo modelo.
- Kilometraje medio.
- Distribución geográfica.
- ...

Punto 8

Licencia

La práctica está orientada para obtener información para particulares sin uso comercial, por este motivo se opta por utilizar la licencia Creative Commons de uso no comercial:

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

Punto 9

Código

El código se encuentra en el repositorio de Github <https://github.com/afcastrom/TCVD-PRA1>.

Se ha realizado un único script denominado **CochesVentasScrap.py**

Punto 10

Dataset

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

<https://doi.org/10.5281/zenodo.3750142>

Punto 11

Trabajo

La práctica y los ficheros solicitadas se presentan en el repositorio de GitHub indicado.

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