Web Scraping

Jonathan Rodriguez Naranjo

8 de noviembre de 2018

Table of Contents

Descripción de la Práctica

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos de una web de venta de vehículos. Para la realización de esta práctica hemos realizado un estudio previo con las siguientes tareas detalladas

- Ver archivo robots.txt:

User-agent: *

Disallow: /wp-admin/

Allow: /wp-admin/admin-ajax.php

- Mapa del sitio web:

URL of sub-sitemap Last modified (GMT)

https://autosroso.com/sitemap-misc.xml 2018-09-11 14:58 https://autosroso.com/sitemap-pt-post-2018-01.xml 2018-01-02 13:42 https://autosroso.com/sitemap-pt-page-2018-08.xml 2018-08-13 16:20 https://autosroso.com/sitemap-pt-page-2018-01.xml 2018-07-02 19:38

URL Priority Change frequency Last modified (GMT)

https://autosroso.com/sobre-nosotros/ 60% Weekly 2018-01-18 19:00 https://autosroso.com/contacto/ 60% Weekly 2018-07-02 14:35 https://autosroso.com/garantias/ 60% Weekly 2018-07-02 13:21 https://autosroso.com/vehiculos/ 60% Weekly 2018-07-02 19:38

- Obtener su tamaño:

8 resultados (0,20 segundos)

- Tecnología usada:

{'web-servers': ['Apache'], 'font-scripts': ['Font Awesome', 'Google Font API'], 'web-frameworks': ['Twitter Bootstrap'], 'ecommerce': ['WooCommerce'], 'cms': ['WordPress'], 'programming-languages': ['PHP'], 'blogs': ['PHP', 'WordPress'], 'javascript-frameworks': ['jQuery']}

- Ver propietario:

{ "domain_name": "AUTOSROSO.COM",

"registrar": "Arsys Internet, S.L. dba NICLINE.COM",

"whois_server": "whois.nicline.com",

Las características del dataset son las siguientes.

1. Título del dataset.

Precios de vehículos de la competencia.

2. Subtítulo del dataset.

Referencia orientativa del precio de venta de los vehículos de la competencia.

3. Imagen identificativa del dataset.

4. Contexto. ¿Cuál es la materia del conjunto de datos?

Este conjunto de datos recoge características y precios de los vehículos. Los datos abarcan el precio de venta, la marca, el modelo y las características. Son vehículos de segunda mano publicitados por empresas que se dedican a la compra-venta de vehículos siendo de utilidad para obtener diferentes tipos de análisis.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

En el conjunto de datos se recogen los siguientes campos:

- Precio: Precio de venta del vehículo.
- Marca: La marca del vehículo.
- Modelo: El modelo del vehículo.
- Potencia: Potencia del vehículo expresado en caballos (cv).
- Versión: Detalle de la versión del modelo.
- Puertas: Número de puertas del vehículo.
- Combustible: Tipo de combustible del vehículo.
- Plazas: Número de plazas del vehículo.
- Co2: Emisiones de Co2 del vehículo expresada gr/km.
- Color: Color exterior vehículo.

- Año: Año de primera matriculación del vehículo.
- Km: Kilómetros recorridos del vehículo.
- Cambio: Tipo de cambio de marchas del vehículo.

El conjunto de datos cambia cada día, con lo cual obliga a automatizar la tarea de raspado para obtener la inserción o eliminación de vehículos a la venta.

La información ha sido recogida mediante web scraping desde "https://autosroso.com/vehiculos/". Esta web carga en un iframe la lista de los vehículos en venta. Los datos se cargan desde "https://publicaciones.carfactory.es/vehicles/" ya que se utiliza un software multi-publicación.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

Se agradece a la empresa autosroso la publicación de los datos los cuales nos han servido para realizar este estudio.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Es interesante obtener los datos de ventas de vehículos en una zona, localidad o provincia. Estos datos sirven para realizar seguimiento de los precios de la competencia, obteniendo desde los sitios web de potenciales competidores información y así comparar precios de determinado modelo con ciertas características. Con esta información se puede tener una ventaja competitiva respecto a las demás empresas. Con dichos datos es posible realizar distintos análisis como son:

- Obtener precios de venta de la competencia para obtener ventaja competitiva. ¿En cuánto venden cierto coche con determinadas características la competencia?
- Conocer precios de venta óptimo de un vehículo por cuenta propia. ¿En cuánto se puede vender mi coche de forma particular?
- Realizar estudios de diferencias de precios en periodos. ¿Cómo se ha depreciado/revalorizado un vehículo?
- Toma de decisiones en el negocio. ¿Es rentable la venta de cierto modelo de vehículo?

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

- Released Under CC BY-NC-SA 4.0 License

Se ha decidido utilizar esta licencia ya que permite:

- Compartir: copiar y redistribuir el material en cualquier medio o formato
- Adaptar: volver a mezclar, transformar y crear a partir del material

Bajo las condiciones siguientes:

- Reconocimiento: Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.
- No Comercial: No puede utilizar el material para una finalidad comercial.
- Compartir Igual: Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la misma licencia que el original.
- No hay restricciones adicionales: No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset

main.py

datascraper.py

10.Dataset: Dataset en formato CSV

pricescar.csv