PRÁCTICA 1

Realización de Web Scraping con Python en la página web de Mil Anuncios y posterior almacenamiento en soporte CSV-Excel

Tipología y ciclo de datos

Contenido

1. Componentes del grupo

David Quiles Gómez

Iván López-Baltasar Benito

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

En la red disponemos de varias webs y portales que contienen gran información relativa a la compra-venta de vehículos de segunda mano, km 0 y de ocasión. Realizando Web Scraping a las principales web de este tipo podríamos obtener información muy rápidamente de gran cantidad de vehículo de cualquier provincia de España faciliando además todas aquellas carácteristicas que suelen buscar los usuarios como Kilometraje, precio, estado, color, etc.

Hemos elegido la pagina Web MilAnuncios.com debido a que contiene gran cantidad de información bien estructurada y responde prácticamente a todas las preguntas que suelen formularse los usuarios. En esta Web, se dispone de información no sólo de particulares sino también de concesionarios que publican semanalmente gran cantidad de información de todos sus vehículos. Además, esta web también nos ofrece información de cualquier parte de España.

3. Definir un título para el dataset. Que sea descriptivo.

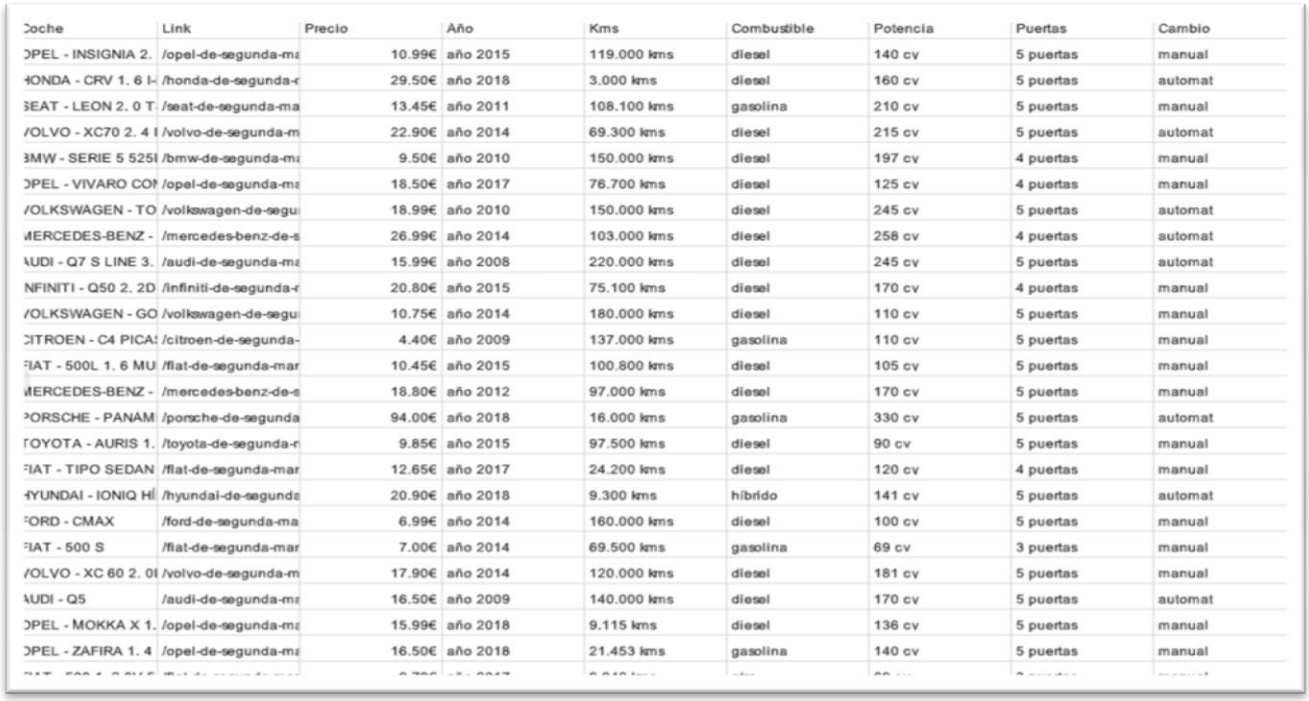
El título elegido para el dataset será "FindMyCar".

Obviamente elegimos un título que pueda dar idea rápida de cuál será su utilidad.

4. Descripción del dataset. Desarrollar una breve descripción de la información aportada.

El Dataset obtenido muestra toda la información que un usuario corriente podría considerar importante a la hora de comprar un vehículo. El dataset se muestra en filas fácilmente manipulable para posteriormente poder ser fitrado, manipulado, almacenado en una BBDD o bien ser tratado con alguna técnica de Minería de datos como Clustering o Clasficación por ejemplo, si los consideramos necesario y nuestro objetivos va más allá que del simple hecho de comprar un coche. Quizás nos interese conocer en qué ciudad se vende más un tipo de coche o en que provincia disponemos de mayor stock, etc.

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



Coche	Link	Precio	Año	Kms	Combustible	Potencia	Puertas	Cambio
OPEL - INSIGNIA 2.	/opel-de-segunda-ma	10.99€	año 2015	119.000 kms	diesel	140 cv	5 puertas	manual
HONDA - CRV 1. 6 I-	/honda-de-segunda-c	29.50€	año 2018	3.000 kms	diesel	160 cv	5 puertas	automat
SEAT - LEON 2. 0 T	/seat-de-segunda-ma	13.45€	año 2011	108.100 kms	gasolina	210 cv	5 puertas	manual
VOLVO - XC70 2. 4 I	/volvo-de-segunda-m	22.90€	año 2014	69.300 kms	diesel	215 cv	5 puertas	automat
BMW - SERIE 5 525I	/bmw-de-segunda-ma	9.50€	año 2010	150.000 kms	diesel	197 cv	4 puertas	manual
OPEL - VIVARO CO	/opel-de-segunda-ma	18.50€	año 2017	76.700 kms	diesel	125 cv	4 puertas	manual
VOLKSWAGEN - TO	/volkswagen-de-segu	18.99€	año 2010	150.000 kms	diesel	245 cv	5 puertas	automat
MERCEDES-BENZ -	/mercedes-benz-de-s	26.99€	año 2014	103.000 kms	diesel	258 cv	4 puertas	automat
AUDI - Q7 S LINE 3.	/audi-de-segunda-ma	15.99€	año 2008	220.000 kms	diesel	245 cv	5 puertas	automat
NFINITI - Q50 2. 2D	/infiniti-de-segunda-c	20.80€	año 2015	75.100 kms	diesel	170 cv	4 puertas	manual
VOLKSWAGEN - GO	/volkswagen-de-segu	10.75€	año 2014	180.000 kms	diesel	110 cv	5 puertas	manual
CITROEN - C4 PICA	/citroen-de-segunda-	4.40€	año 2009	137.000 kms	gasolina	110 cv	5 puertas	manual
FIAT - 500L 1. 6 MU	/fiat-de-segunda-mar	10.45€	año 2015	100.800 kms	diesel	105 cv	5 puertas	manual
MERCEDES-BENZ -	/mercedes-benz-de-s	18.80€	año 2012	97.000 kms	diesel	170 cv	5 puertas	manual
PORSCHE - PANAM	/porsche-de-segunda	94.00€	año 2018	16.000 kms	gasolina	330 cv	5 puertas	automat
TOYOTA - AURIS 1.	/toyota-de-segunda-c	9.85€	año 2015	97.500 kms	diesel	90 cv	5 puertas	manual
FIAT - TIPO SEDAN	/fiat-de-segunda-mar	12.65€	año 2017	24.200 kms	diesel	120 cv	4 puertas	manual
HYUNDAI - IONIQ HÍ	/hyundai-de-segunda	20.90€	año 2018	9.300 kms	híbrido	141 cv	5 puertas	automat
FORD - CMAX	/ford-de-segunda-ma	6.99€	año 2014	160.000 kms	diesel	100 cv	5 puertas	manual
FIAT - 500 S	/fiat-de-segunda-mar	7.00€	año 2014	69.500 kms	gasolina	69 cv	3 puertas	manual
VOLVO - XC 60 2. 0I	/volvo-de-segunda-m	17.90€	año 2014	120.000 kms	diesel	181 cv	5 puertas	manual
AUDI - Q5	/audi-de-segunda-ma	16.50€	año 2009	140.000 kms	diesel	170 cv	5 puertas	automat
OPEL - MOKKA X 1.	/opel-de-segunda-ma	15.99€	año 2018	9.115 kms	diesel	136 cv	5 puertas	manual
OPEL - ZAFIRA 1. 4	/opel-de-segunda-ma	16.50€	año 2018	21.453 kms	gasolina	140 cv	5 puertas	manual

6. Contenido. Explicar los campos del Dataset Obtenido.

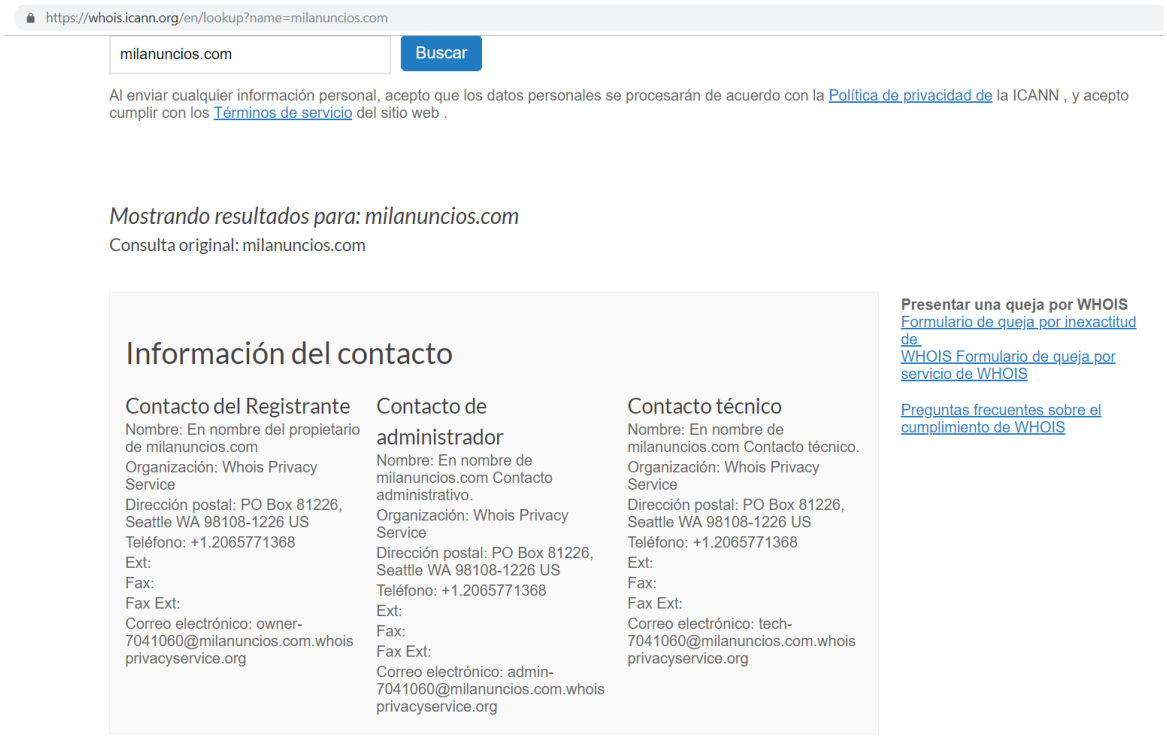
El dataset obtenido tiene la siguiente información:

- MODELO → Modelo y marca del vehículo.
- LINK → El link para acceder directamente a la web donde está todo el contenido.
- PRECIOS → El precio del vehículo
- AÑO MATRICULACIÓN.
- KILOMETROS
- TIPO DE COMBUSTILES → Gasoil, diesel, etc.
- POTENCIA → en CV
- PUERTAS → 3 o 5 Puertas
- CAMBIO → Manual o automático.

7. Agradecimientos. Presentar al propietario del conjunto de datos.

Como no podía ser de otra manera, agradecemos a todo el equipo de milanuncios.com su extraordinario trabajo. Acceder a su web a sido sencillo y sus datos han sido fácilmente manejables.

El propietario al que debemos el éxito de nuestro trabajo es:



Los propietarios de la web no están identificados. Hemos utilizado la herramientas whois para intentar averiguar quien son los propietarios pero su identidad está oculta.

8. Inspiración. Explique por qué es interesante este conjunto de datos y que preguntas puede responder.

Una de las inversiones más complejas y difíciles que tiene cualquier persona es la adquisición de un vehículo. La búsqueda de uno que cubra todas nuestras necesidades y expectativas suele ser cuestión de semanas o incluso meses, visitando diferentes Webs de coches de ocasión, concesionarios, mercado particular, etc.

El proyecto puede ser interesante tanto del lado del comprador como por parte del vendedor

Por parte del comprador, a través de un proyecto de Web Scraping podemos obtener información precisa de todas esas Webs de manera rápida y concreta y que sin duda nos ahorrará decenas de horas de busqueda por las diversas webs que se dedican a la compraventa de vehículos.

No obstante creemos que la información obtenida es mucho más interesante desde el punto de vista del vendedor puesto que con los obtenidos y obteniendo los datos de coches vendidos podríamos establecer relaciones entre las ventas y las compras en las diferentes provincias de España. Podríamos contestar, como ejemplos, a las siguientes preguntas

- Que coches se venden más en cada una de las provincias
- Qué tipo de coches se venden más en cada una de las provincias (usados, nuevos, etc)
- Que rango de precio tienen los coches vendidos (en cualquier provincia)
- En qué fechas se venden más coches
- Duración media de los stock de los modelos de los diferentes marcas, nuevamente por provincias, etc, etc.
- Que concesionarios tienen precios más competitivos
- Precios medio de los modelos vendidos en cada provincia para comprender si (p.ej.) si la marca Peugeot tiene precios medios más económicos en una provincia que en otra.

Hay inmobiliarias que están comenzando a incorporar a Científicos de Datos para establecer relaciones entre los bienes inmobiliarios y los compradores, de manera que la inmobiliaria puede saber si un piso de 120 m2 con un precio de 145.000 en Getafe será vendido antes de 60 días. La contestación a todas estas preguntas nos podría llevar a sacar conclusiones de si nos interesa adquirir, para después vender, un bien inmobiliario en una determinada ciudad.

De manera análoga, este interesante ejercicio, podría trasladarse a los coches, camiones, etc para tomar decisiones relevantes y optimizar nuestros recursos económicos.

Por supuesto, disponer de este conjunto de datos nos permitirá utilizar técnicas de minería de datos capaces de responder a todas estas preguntas. Este dataset y debido a su claridad es fácilmente manipulable, escalable para poder crear modelos útiles.

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

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

Cabe recordar inicialmente que estas licencias hacen referencia a las licencias Creative Commons y que cada una de ellas tiene posibilidades para configurarlas y que permite a los autores de los dataset poder decidir de qué manera va a circular en internet su trabajo. Dependiendo de esta configuración, los usuarios podrán publicar, citar, reproducir y crear obras derivadas utilizando el trabajo publicado generalmente bajo ciertas restricciones.

Explicamos rápidamente algunas siglas (llamados módulos oficialmente) de estas licencias:

CC→ Creative Commons (abreviatura)
BY → requiere la referencia al autor original.
SA → permite obras derivadas bajo la misma licencia o similar.
NC → obliga a que la obra no sea utilizada con fines comerciales.
ND → no permite modificar la obra de ninguna manera.
CCO → Sin derechos reservados.

Estos módulos asimismo se combinan entre si para formar las seis licencias de Creative Commons:

- Atribución (CC – BY)
- Atribución Compartir Igual (CC-BY-SA)
- Atribución No Derivadas (CC-BY-ND)
- Atribución No Comercial (CC-BY-NC)
- Atribución No comercial Compartir Igual (CC-BY-NC-SA)
- Atribución No comercial No derivadas (CC-BY-NC-ND)

Todas las licencias permiten el derecho fundamental de redistribuir la obra con fines NO comerciales y sin modificaciones, aunque las opciones NC y ND hacen que la obra no sea de libre acuerdo para tal redistribución.

Una vez definidas y recordadas las posibilidades, nuestra licencia tendrá las siguientes características:

- BY, puesto que requerirá al autor original
- NC, puesto que no podremos utilizarla para fines comerciales. Entendamos que milanuncios recibe dinero debido a la publicidad que insertan en su web y que se presupone que ven las personas que la visitan y que obviamente la información que obtenemos de su web es información en propiedad y protegida posiblemente por la LPD.
- ND, puesto que no podremos modificar los registros bajo ningún concepto

Así pues, el trabajo podría tener la licencia más abajo indicada y que no está indicada en las opciones del enunciado:

- Atribución NO comercial NO derivadas (CC-BY-NC-ND)

Una licencia alternativa sería

- Released Under CC BY-NC-SA 4.0 License,

Que permitiría derivar el trabajo bajo la misma licencia y también sin fines comerciales.

Fuente: https://es.m.wikipedia.org/wiki/Creative_Commons

Las licencias Creative Commons están compuestas por cuatro módulos de condiciones:

- *Attribution/Atribución (BY), requiere la referencia al autor original.*
- *Share Alike/Compartir Igual (SA), permite obras derivadas bajo la misma licencia o similar (posterior u otra versión por estar en distinta jurisdicción).*
- *Non-Commercial/No Comercial (NC), obliga a que la obra no sea utilizada con fines comerciales.*
- *No Derivative Works/No Derivadas (ND), no permite modificar la obra de ninguna manera.*

10. Contribuciones

Contribuciones	Firma
Investigación previa	David Quiles / Iván López-Baltasar
Redacción de las respuestas	David Quiles / Iván López-Baltasar
Desarrollo código	David Quiles / Iván López-Baltasar