Pràctica 1: Web scraping

Membres del equip: Alonso Lopez Vicente i Aitor Ferrus Blasco

Usuaris UOC: alopezvic i aferrus

Estudis que cursa (Màster o Grau): Màster en Ciència de Dades

Índex

Context

Aquesta pràctica s'ha realitzat sota el context de l'assignatura Tipologia i cicle de vida de les dades, pertanyent al Màster en Ciència de Dades de la Universitat Oberta de Catalunya. En ella, s'apliquen tècniques de web scraping mitjançant el llenguatge de programació Python per a extreure, en una data concreta, els diversos models de cotxes que ofereix a Espanya l'empresa d'automoció Suzuki. En concret, les dades capturades són les diferents versions de cada model, l'acabat i el seu preu. Les dades s'han extret de la pàgina web https://auto.suzuki.es/ .

Per a l'extracció de les dades d'aquesta pàgina web, s'ha tingut en compte l'arxiu robots.txt (https://auto.suzuki.es/robots.txt), el qual en aquest cas ens indica que no hi ha cap restricció. Encara que l'arxiu haguera esmentat alguna restricció hem de recordar que aquestes són només suggeriments i mai una obligació. En quant el mapa web per aquesta pàgina web és inexistent.

Títol

El títol del nostre data set es **Models de cotxes Suzuki a Espanya**

Descripció

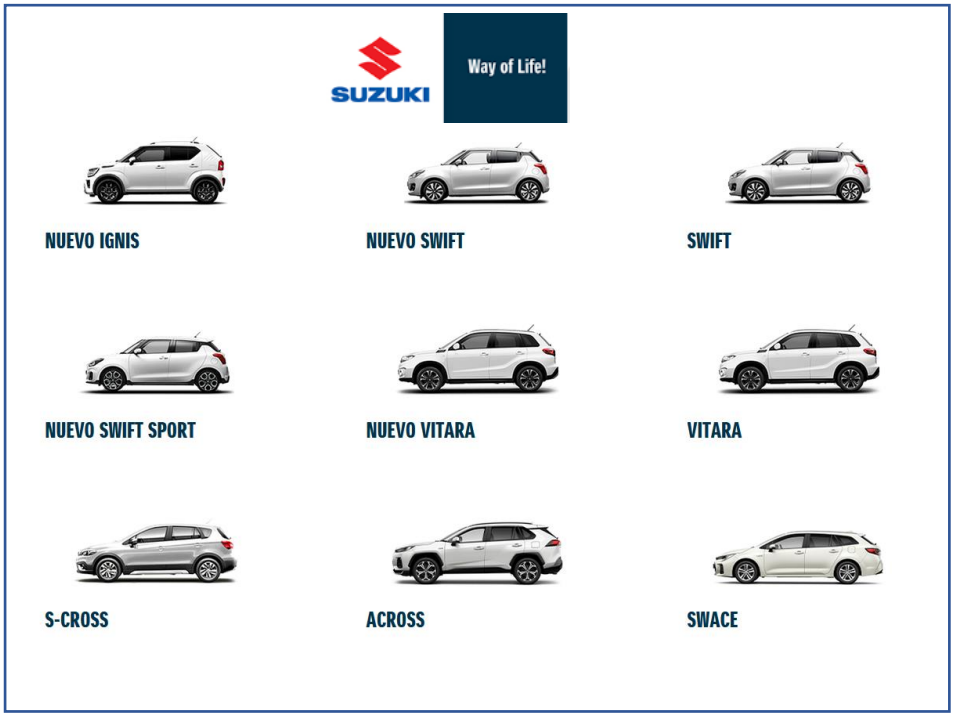
Una vegada examinada la web, hem localitzat les dades a la pàgina https://auto.suzuki.es/precios

A continuació, hem analitzat l'estructura per tal de localitzar els elements que cercàvem, i que es trobaven en taules, tal com s'explica en el codi Python.

El conjunt de dades extretes com a part d'aquesta pràctica conté el dia de l'extracció de les dades, el nom, la versió, l'acabat i el preu dels diferents models de cotxes que la pàgina web https://auto.suzuki.es/ oferta a Espanya.

Per exemple: **02/11/2020 - Nuevo Ignis - 1.2L Mild Hybrid - GLE - 14575 €**

Representació gràfica



Contingut

Per a cada model de cotxe es recullen les següents característiques:

- Nom i versió: El nom i la versió del cotxe
- Acabat: El conjunt de complements. Per exemple GLE,GLX o SPORT
- Preu: El preu del cotxe, o alguna etiqueta promocional
- Data d'extracció: el dia en que s'han capturat les dades, en format dd/mm/aaaa

Els autors de la pagina web https://auto.suzuki.es/precios no guarden la informació dels preus dels cotxes en el passat així que tant sols es pot accedir a la informació del present. Aquestes dades son utilitzades a les pagines web de tots els concessionaris Suzuki d'Espanya.

Agraïments

Les dades han sigut recol·lectades de la pàgina web https://auto.suzuki.es/ . Per això, s'ha utilitzat el llenguatge de programació Python i les tècniques de Web Scraping per a extreure la informació que es trobava a les pagines HTML.

Suzuki http://www.suzuki.com/ és una companyia japonesa d'abast mundial que fabrica diferents tipus de vehicles. Té tres grans divisions:

- Motocicletes, scooters i ATV (quads)
- Motors per embarcacions
- Automòbils

En aquest treball ens hem centrat en la divisió d'automòbils i concretament en la delegació a Espanya. Comparant amb les webs d'altres indrets, hem vist que Suzuki ofereix en cada país models diferents, amb uns acabat específics, adequats a la cultura i els gustos del consumidor del país. Això determina les preferències respecte els models oferts per part dels compradors.

Una de les característiques pròpies de Suzuki és que té una política de preus transparent. Això vol dir, que publica els preus dels seus vehicles, a diferència d'altres marques en que els preus no són públics i cal fer una petició per rebre una oferta amb el preu del model concret sol·licitat.

Inspiració

Creiem que la informació capturada pot ajudar a fer un seguiment dels preus i els models que ofereix Suzuki per un particular que estigui interessat en adquirir un vehicle. De fet es podria fer amb d'altres marques per tal de fer comparatives, encara que, com hem assenyalat abans, moltes no ofereixen els preus públicament.

També podria ser recol·lectada amb altres objectius diferents a l'acadèmic o personal. Per exemple, podríem plantejar el següent cas hipotètic: La Fiat, una competidora directa de Suzuki, busca extreure aquestes dades de forma automàtica de la web per tal de decidir quins preus aplicar als seus propis models i així oferir models de cotxes similars al mateix preu o inferior al que ofereix Suzuki a Espanya.

Creiem que aquest tipus de dada és molt interessant, ja que està relacionada amb un mercat de productes, els cotxes, que no es compren diàriament, i pot resultar difícil per als compradors informar-se o identificar les pujades i baixades dels preus. El codi creat podria ser utilitzat per a, de manera automàtica, extreure les dades, dia a dia, durant tot un any i realitzar després un anàlisi de la variació dels preus de cada model. Això podria ser interessant també, tal com hem comentat. per a les empreses competidores de Suzuki.

Llicència

La llicencia escollida per a la publicació d'aquest conjunt de dades es **CC BY-SA 4.0 License.** Aquesta permet el següent:

Compartir — L'usuari pot copiar i redistribuir el material en qualsevol medi o format.

Adaptar — L'usuari pot remesclar, transformar, i utilitzar al material per a qualsevol propòsit, fins i tot comercialment.

Sempre i quan es compleixen les següents condicions:

Atribució — L'usuari ha de donar crèdit apropiat, proporcionar un nexe a la llicència, i indicar si algun canvi va a ser realitzat. Lo anterior pot ser fet de qualsevol manera raonable, però no en una manera que suggereixi que el llicenciador aprova a l'usuari o el seu ús.

Compartir — Si l'usuari remescla, transforma o utilitza al material per a qualsevol propòsit, ell/ella deu distribuir les seus contribucions sota la mateixa llicència com el primigeni.

Cap restricció addicional — L'usuari no pot aplicar termes legals o mesures tecnològiques que legalment restringeixen altres de fer qualsevol cosa que els permisos de llicència permetin.

Donat que estem fent un treball acadèmic, aquest llicència dona llibertat per utilitzar el codi, sempre que no sigui amb finalitats comercials, i que es reconegui els autors. Així mateix, l'obra resultant ha d'estar sotmesa a la mateixa llicència que l'original.

Codi

Enllaç per accedir al repositori en GitHub: https://github.com/aitorf94/Web-scraping .

Dataset

Enllaç per accedir al data set en format CSV a Zenodo: **https://doi.org/10.5281/zenodo.4141952**

Contribucions

Contribucions	Signa
Recerca prèvia	Alonso Lopez Vicente i Aitor Ferrus Blasco
Redacció de les respostes	Alonso Lopez Vicente i Aitor Ferrus Blasco
Desenvolupament codi	Alonso Lopez Vicente i Aitor Ferrus Blasco