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): Máster en Ciència de Dades

# Context (Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.)

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 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/ .

# Títol (Definir un títol pel dataset. Triar un títol que sigui descriptiu.)

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

Suggeriments

1. **AutosSuzukiScraper**
2. **CotxesSuzukiData**
3. **SusukiCarsScraper**
4. **Cotxes Suzuki Espanya**
5. **Models de cotxes Suzuki a Espanya**

# Descripció (Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).)

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 - 14.575 €**

# Contingut (Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.)

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 (Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).)

*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. Vols dir de posar això? Potser millor un cosa com ara...*

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...

**Crec que seria millor explicar una mica què és Suzuki que no les altres webs que ens hem plantejat... Podem posar enllaços a la web mare... Ens podem enrotllar més o menys...**

*No estic segur si en aquest aparat deguem fer referencia a les pagines web que ens hem plantejat prèviament com:*

*[https://www.tmb.cat](https://www.tmb.cat" \t "_blank)*

*[https://www.bsmsa.cat/](https://www.bsmsa.cat/" \t "_blank)*

*[https://www.comuniazo.com](https://www.comuniazo.com" \t "_blank)*

*I quin tipus de cites de recerca podem afegir.*

**Si vols posar quelcom de recerca, podem cercar altres companyies de cotxes: Fiat, per exemple...**

# Inspiració (Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.)

Creiem que aquesta informació 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.

També podria ser recol·lectada amb altres objectius diferents a l’acadèmic. 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 que Suzuki a Espanya.

Aquest tipus de data és molt interessant, ja que està relacionada amb un mercat de productes (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, extraure 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 per a les empreses competidores de Suzuki.

# Llicència (Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció)

<https://creativecommons.org/publicdomain/zero/1.0/deed.es>

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

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.

Aquesta és la que més m’agrada. Ja em diràs que opines...

<https://creativecommons.org/licenses/by-sa/4.0/>

<https://opendatacommons.org/licenses/odbl/>

# Codi (Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.)

Ho deixem pel final, oi?

# Dataset (Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.)

Podem vincular l’alta a Zenodo amb l’usuari de GitHub. A comentar...

# 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 |