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 així, la data , els diversos models de cotxes, les diferents versions de cada model i el seu preu 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 --------------

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 conte el dia de l'extracció de les dades, el nom, la versió, l'acabat i el preu dels diferents 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: El nom del cotxe
* Versió: La versió del cotxe.
* Acabat: El acabat GLE,GLX o SPORT
* Preu: El preu del cotxe
* Data de extracció: Dia de extracció de les dades anteriors 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 a pagina web <https://auto.suzuki.es/> . Per això, s’ha utilitzat el llenguatge de programació Python y les tècniques de Web Scraping per a extraure la informació que es trobava a les pagines HTML.

......

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.bsmsa.cat/>

<https://www.comuniazo.com>

I quin tipus de cites de recerca podem afegir.

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

Aquesta informació també podria ser recol·lectada amb altres objectius diferents al acadèmic. Un exemple seria els següent hipotètic cas: La Fiat una competidora directa de Suzuki busca extreure aquestes dades de forma automàtica de la web per així 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 es refereix a un mercat amb productes (cotxes) que no son productes que es compren i comparen diàriament, i pot resultar difícil per als compradors informar-se o identificar les pujades i baixades dels preus. El codic creat, podria ser utilitzat per a automàticament 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/>

<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.)

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

# 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 |