Universitat Oberta de Catalunya

Deadline: 09/11/2020

# Tipologia i cicle de vida de les dades

*Pràctica 1 – Web scraping*


Albert Gil Devesa
Usuari GitHub: **agildeve**

Aleix Borrella Colomé
Usuari GitHub: **iAleix**

***Màster Universitari de Ciència de Dades***

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

Per a la realització d'aquesta pràctica hem considerat la opció de fer web scraping sobre un supermercat online. Creiem que pot ser un bon exemple de l'ús d'aquesta eina ja que ofereix moltes possibilitats. Una d'elles podria ser que, per exemple, una empresa demani un anàlisi de la competència per comparar preus i ofertes, així com la seva evolució al llarg del temps. Per altra banda, també pot ser útil per realitzar un estudi intern creuant aquesta informació amb la informació de ventes, per veure quines ofertes són més atractives, quins productes tenen major demanda amb/sense promoció, etc.

Respecte als diferents supermercats online que existeixen, hem escollit el web del Bonpreu-Esclat ja que aquest proporciona una bona quantitat de productes ben classificats per categories i no hi prohibeix l'ús de bots, tal i com es pot veure en el seu arxiu *"robots.txt"*, on només s'exclouen tres directoris:

https://www.bonpreuesclat.cat/robots.txt

```
User-Agent: *
Disallow: /group*/
Disallow: /c/portal/
Disallow: /*?p_p_id=3
Sitemap: https://www.bonpreuesclat.cat/sitemap.xml
```

**2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.**

Com hem comentat, en aquesta pràctica ens hem centrat en realitzar un scraping de la informació continguda en el supermercat online que ofereix la marca **Bonpreu-Esclat.** Així doncs, hem cregut oportú donar per títol al dataset el nom de:

**Bonpreu-Esclat Online Supermarket Products**

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

En el marc d'aquesta pràctica s'han capturat tots els productes que s'ofereixen al supermercat online de Bonpreu-Esclat. Durant aquesta captura, hem obtingut informació com: el nom del producte, el seu preu, la quantitat de producte, el preu per unitat de referència, cinc categories de classificació, si el producte es troba en oferta o no, quina promoció té, l'enllaç url del producte i la data en la que s'han obtingut els valors.

A més, el dataset que s'obté amb l'execució del codi és un arxiu .csv que porta per nom "yyyymmdd_bp_dataset.csv", on "yyyymmdd" fa referència a la data del dia en que s'ha pres la mostra. Posteriorment, es poden concatenar els diferents arxius .csv per poder respondre preguntes de caire temporal (duració de les promocions, evolució dels preus, etc), la qual cosa explicarem més en detall en l'apartat 5.

## 4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment

Una manera il·lustrativa de representar el dataset per tenir una primera impressió gràfica sobre les dades que conté podria ser la següent:


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

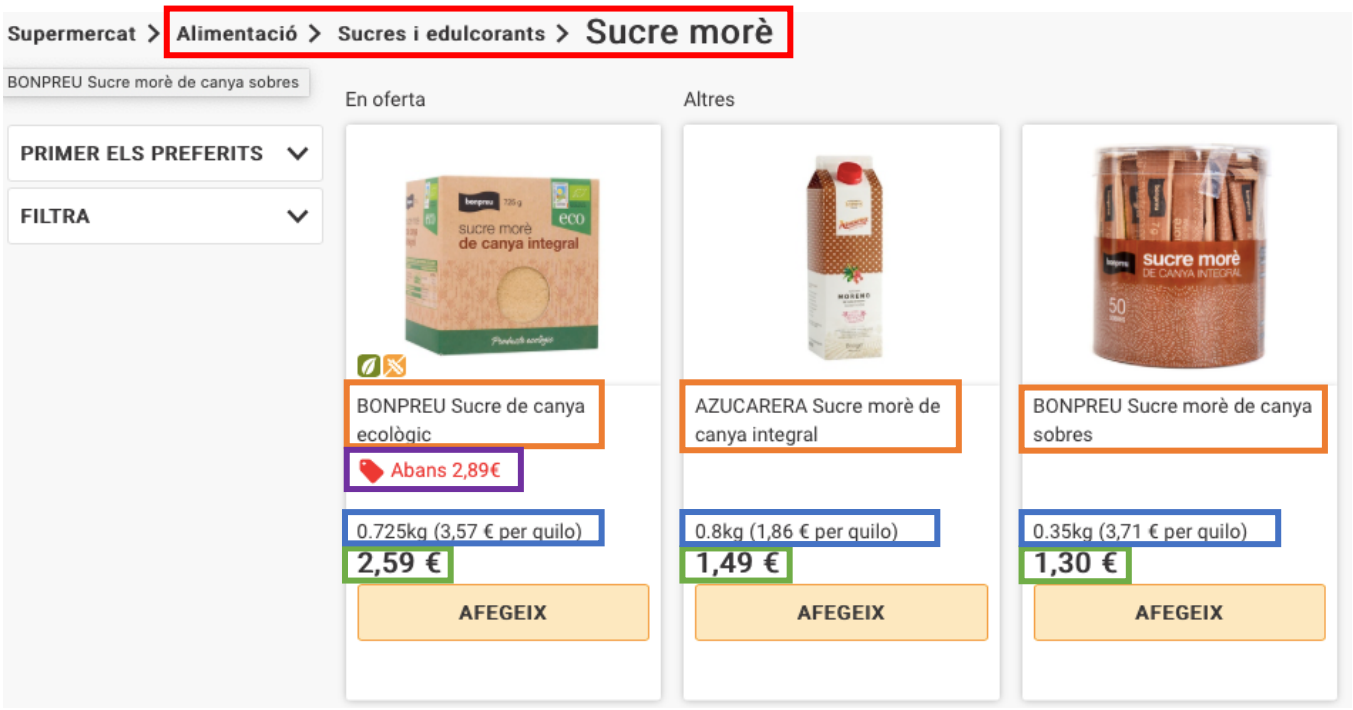
El nostre dataset conté 13 camps, els quals detallem a continuació:

1. **Categoria_1:** Primer nivell de classificació de productes.
2. **Categoria_2:** Segon nivell de classificació de productes.
3. **Categoria_3:** Tercer nivell de classificació de productes.
4. **Categoria_4:** Quart nivell de classificació de productes.
5. **Categoria_5:** Cinquè nivell de classificació de productes.
6. **Nom:** Nom amb el que s'oferta el producte.
7. **Preu:** Preu amb el que s'oferta el producte.
8. **Quantitat:** Quantitat de producte que conté l'article.
9. **Preu unitari:** Preu del producte, respecte a la unitat de quantitat referència.
10. **Oferta:** Si el producte està subjecte a alguna oferta o promoció.
11. **Promoció:** Descripció de l'oferta o promoció, si aplica.
12. **URL:** URL del producte.
13. **Data:** Data en la qual s'ha capturat la informació.

Com hem comentat anteriorment, el scraping que hem dissenyat genera un arxiu .csv cada dia, així que per poder obtenir dades amb evolució temporal hem realitzat una captura diària durant la setmana laboral del 02 al 06 de Novembre de 2020, obtenint 5 arxius .csv diferents.

Posteriorment, aquests 5 arxius .csv obtinguts que porten per nom “yyyymmdd_bp_dataset.csv” els hem concatenat en un sol arxiu .csv anomenat “BP_dataset.csv”, el qual conté la evolució de la informació corresponent a tots els productes que ofereix el supermercat online al llarg de la setmana.

Per obtenir els diferents camps que formen cada dataset, el sistema de scraping que hem dissenyat es basa principalment en el següent esquema:




| Categoria_1 | Categoria_2 | Categoria_3 | Categoria_4 | Categoria_5 | Nom | Preu | Quantita | Preu unitari | Oferta | Promocio | URL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Alimentació | Sucres i edulcorants | Sucre morè | - | - | BONPREU Sucre de canya ecològic | 2,59 € | 0.725kg | (3,57 € per quilo) | Si | Abans 2,89€ | https://www.co |
| Alimentació | Sucres i edulcorants | Sucre morè | - | - | AZUCARERA Sucre morè de canya integral | 1,49 € | 0.8kg | (1,86 € per quilo) | No | - | https://www.co |
| Alimentació | Sucres i edulcorants | Sucre morè | - | - | BONPREU Sucre morè de canya sobres | 1,30 € | 0.35kg | (3,71 € per quilo) | No | - | https://www.co |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

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

Per obtenir el nostre dataset, hem extret les dades a partir del supermercat online que ofereix la marca Bonpreu-Esclat, així que desconeixem la existència de si hi ha algun propietari del conjunt de les dades. En tot cas, podem fer ús de les tècniques vistes en els apunts per veure informació més en detall sobre el lloc web:

```
import whois

print(whois.whois('https://www.bonpreuesclat.cat'))
```


```
{
  "domain_name": "bonpreuesclat.cat",
  "registrar": "UBILIBET S.L.",
  "whois_server": null,
  "referral_url": null,
  "updated_date": "2020-10-18 12:15:05.888000",
  "creation_date": "2015-09-03 11:25:38.287000",
  "expiration_date": "2021-09-03 11:25:38.287000",
  "name_servers": [
    "ns3.ascio.com",
    "ns4.ascio.com",
    "ns1.ascio.com",
    "ns2.ascio.com"
  ],
  "status": "ok https://icann.org/epp#ok",
  "emails": "registrar@ubilibet.com",
  "dnssec": "unsigned",
  "name": null,
  "org": "Bon Preu, SAU",
  "address": null,
  "city": null,
  "state": "Barcelona",
  "zipcode": null,
  "country": "ES"
}
```


Podem veure que el lloc web està registrat per *"UBILIBET S.L."*, també ho està en el "*ICANN*", i és propietat de l'organització *"Bon Preu, SAU*" , localitzada a Barcelona. Així doncs, els agraïments estan destinats principalment a aquesta organització que ofereix el web per al supermercat online, a partir del qual hem obtingut les dades.

També comentar que per extreure la informació del lloc web s'ha utilitzat Python com a llenguatge de programació, així com les eines de web scraping presentades i explicades en els apunts de l'assignatura, les quals ens han permès obtenir la informació continguda en les pàgines HTML. Per tant, creiem important també agrair els apunts proporcionats en l'aula que ens han permès realitzar la pràctica, concretament:

- *El llenguatge Python - David Masip Rodó*
- *Web Scraping - Laia Subirats Maté i Mireia Calvo González*
- *Web Scraping with Python – Chapter 2: Scraping the data – Lawson, R*
- *Automated Data Collection with R – S. Munzert, C. Rubba, P. Meibner i D. Nyhuis*

Per últim, només comentar que prèviament a l'extracció de les dades hem fet algunes recerques que ens han ajudat a tenir una primera idea de com orientar la pràctica, concretament:

- Web scraping for food price research - Judith Hille
- https://towardsdatascience.com/how-to-scrape-google-shopping-prices-with-web-data-extraction-5a0a9b92406f
- https://www.youtube.com/watch?v=ng2o98k983k&t=2321s

**7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.**

Tothom necessita realitzar la compra per abastir-se d'aliments i altres consumibles de primera necessitat. En els supermercats trobem una gran quantitat d'aquests productes i és molt probable que en el futur més i més cadenes desenvolupin supermercats online com el que hem emprat en el nostre estudi. Podem enfocar, doncs, l'anàlisi d'aquest dataset des de diferents interessos:

- **Com a clients:** Podem analitzar quina varietat de productes hi ha, quin estalvi obtenim degut a les promocions, quina és la diferència de preu entre productes que comprem a un altre supermercat i els preus de Bonpreu-Esclat.

- **Com a Bonpreu-Esclat:** Creuant aquestes dades amb altres d'internes podem veure com d'eficaces són les promocions, quin és el temps òptim de durada d'una promoció, si la varietat de productes és adequada (pot ser que hi hagi massa oferta d'un tipus concret de producte), quines preferències tenen els clients en quant a marques-cost, etc.

- **Com a un anàlisi de la competència:** Quina durada tenen les promocions, quina diferència de preus hi ha entre ambdós supermercats, quina varietat d'oferta té Bonpreu-Esclat envers el nostre supermercat o quant d'econòmiques són les promocions, entre d'altres.

En el nostre cas, ens hem plantejat respondre les següents preguntes:

- Quantes promocions hi ha?
- Com varien les promocions al llarg de la setmana?
- Quin estalvi (en percentatge) mitjà obtenim amb les promocions?
- Quins articles tenen el major descompte? De quant és?
- Quins articles tenen el menor descompte? De quant és?
- Quants articles hi ha per categoria?
- Hi ha articles que han variat el seu preu?
- Com evoluciona el preu dels articles al llarg de la setmana?
- Hi ha noves ofertes o promocions? Se'n ha retirat alguna?

## 8. Llicència. Seleccionar una llicència pel dataset resultant i explicar el motiu de la seva selecció:

En l'enunciat d'aquesta pràctica s'anomenen diferents llicències a triar pel nostre dataset resultant. No obstant això, abans d'escollir-la, cal entendre que implica cadascuna:

- **Released Under CC0 Public Domain License:** Aquesta llicència podríem dir que és la menys restrictiva de totes ja que fa referència a *"No rights reserved"*, és a dir, que tothom pot fer ús de la base de dades sense que hi hagi cap tipus de restricció de copyright.

- **Released Under CC BY-NC-SA 4.0 License:** Aquesta llicència, en canvi, és de les més restrictives ja que fa referència a *"Atribution-NonCommercial-ShareAlike",* és a dir, que bàsicament el nostre dataset només podrà ser compartit i adaptat. Així doncs, el dataset serà atribuït al propietari de la llicència (al qual caldrà citar si es comparteix), no podrà ser utilitzat amb finalitats comercials i, si és adaptat o distribuït, caldrà fer-ho sota la mateixa llicència.

- **Released Under CC BY-SA 4.0 License:** Aquesta llicència és molt similar a l'anterior però en aquest cas només fa referència a "*Atribution-ShareAlike",* és a dir, que amb les mateixes restriccions que en el cas anterior el dataset podrà ser compartit i adaptat, però en aquest cas si que se li podran donar finalitats comercials.

- **Database released under Open Database License, individual contents under Database Contents License:** Aquesta llicència és semblant a la CC BY-SA 4.0 però en aquest cas també es fa referència a "*Atribution-ShareAlike-KeepOpen"*, és a dir, que el dataset podrà ser compartit i adaptat sempre fent referència al propietari, sota la mateixa llicència i, en aquest cas, s'afegeix que el dataset resultant sempre haurà de ser en format de dades obertes.

- **Other:** També existeixen altres tipus de llicències com podrien ser la CC BY-ND o la CC BY-NC-ND, entre d'altres.

- **Unknown License:** Aquells datasets dels quals se'n desconeix la llicència.

En el nostre cas, donat que es tracta d'un dataset obtingut a partir d'una pràctica referent a uns estudis universitaris, en principi no se li donarà cap ús comercial. També és veritat que no ens agradaria compartir el dataset sense cap tipus de llicència (CC0) ja que ens ha suposat un esforç obtenir-lo, així que si les dades són compartides, caldria citar-nos. A més, ens agradaria que si les dades que hem obtingut són usades, aquestes sempre acabin sent compartides amb la mateixa llicència que li donarem per tal de poder ser reutilitzades.

Així doncs, analitzant els nostres criteris, creiem que la llicència més adient per al nostre cas d'estudi seria la **"CC BY-NC-SA 4.0 License".**

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

El codi que hem fet servir per a generar el dataset està desenvolupat amb llenguatge Python i està format per 4 arxius .py diferents: un primer codi es dedica a entrar a tots els webs que té el supermercat online i adquireix les urls dels productes que es troben en les últimes subcategories de classificació possibles per tal d'evitar duplicats; per altra banda, un segon codi desenvolupa una funció que, donada una url, fa un web scraping de tots els camps comentats anteriorment que aquesta conté i els va guardant en el dataset resultant; un tercer codi simplement crida els dos codis anteriors perquè s'executin; i un quart codi s'encarrega de concatenar els 5 arxius .csv que s'han obtingut per cada dia de la setmana laboral en un sol dataset, que serà el resultant.

Per veure-ho de forma més esquemàtica, els 4 codis que hem creat es resumeixen en:

- ***get_urls.py*:** Obté un llistat de totes les urls del supermercat online.
- ***scraper.py*:** Realitza el web scraping per a cada una de les urls anteriors.
- ***main.py*:** Crida els codis anteriors ("*get_urls.py*" i "*scraper.py*").
- ***concat.py*:** Concatena els 5 arxius .csv obtinguts cada dia en una dataset conjunt.

Ens agradaria comentar que podríem haver generat un sol arxiu .py que contingués tota la informació d'aquests 4 arxius comentats, però creiem que de cara a la detecció d'errors és molt més pràctic treballar amb cada part per separat.

Tant els diferents codis utilitzats, com els 5 arxius .csv obtinguts al llarg de la setmana, com el dataset resultant de la concatenació dels .csv es poden trobar en el nostre repositori de GitHub, concretament en el següent enllaç:

***https://github.com/iAleix/PRA1_Web-Scraping***

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

Tal i com es demana, també hem publicat el dataset resultant (en aquest cas hem optat per només publicar el dataset final i no els intermedis) en format .csv a Zenodo, obtenint la DOI corresponent. Concretament es troba al següent enllaç:

***https://zenodo.org/record/4263423#.X6hzPy8rxQI***

***(DOI:** 10.5281/zenodo.4263423)*

***Taula de contribucions al treball***

| Contribucions | Signatura |
|---|---|
| Recerca prèvia | ABC / AGD |
| Redacció de les respostes: | |
| 1. Context | ABC / AGD |
| 2. Definir un títol pel dataset | ABC / AGD |
| 3. Descripció del dataset | ABC / AGD |
| 4. Representació gràfica | ABC / AGD |
| 5. Contingut | ABC / AGD |
| 6. Agraïments | ABC / AGD |
| 7. Inspiració | ABC / AGD |
| 8. Llicència | ABC / AGD |
| 9. Codi | ABC / AGD |
| 10. Dataset | ABC / AGD |
| Desenvolupament del codi | ABC / AGD |