Tipologia i cicle de vida de les dades – PRA1

Estudiants: Jordi Dil Giró i Carlota Font Castell

Enllaç al Github: https://github.com/cfont03/PRA1-Abacus

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

El model de negoci tradicional ha experimentat un gran canvi en els darrers anys. En el seu moment i, encara avui dia, el petit negoci ha de fer front a les grans superfícies. Actualment, però, cal també afrontar la competència online.
L'empresa "La llar del llibre" ha notat una forta baixada de ventes en llibres en els darrers 18 mesos. Durant aquest temps, de forma poc estructurada, l'empresa ha dut a terme diverses metodologies per tal de recuperar-ne les ventes: abaixar preus, incrementar la promoció online, millorar els temps d'entrega en la venta online, entre d'altres.
Després de veure que aquestes no han sorgit quasi efecte, l'empresa ha decidit focalitzar-se en la seva competència més directe i més forta online, l'empresa Abacus Cooperativa, especialment després d'haver llegit la següent notícia: https://eshowmagazine.com/ultimas-noticias/abacus-incrementa-sus-ventas-online-un-245-en-2018/
Amb l'objectiu de recuperar-ne el mercat, l'empresa s'ha adreçat a dos estudiants de la Universitat Oberta de Catalunya per tal de fer una anàlisi sobre els llibres venuts online en la pàgina web de l'Abacus Coop.

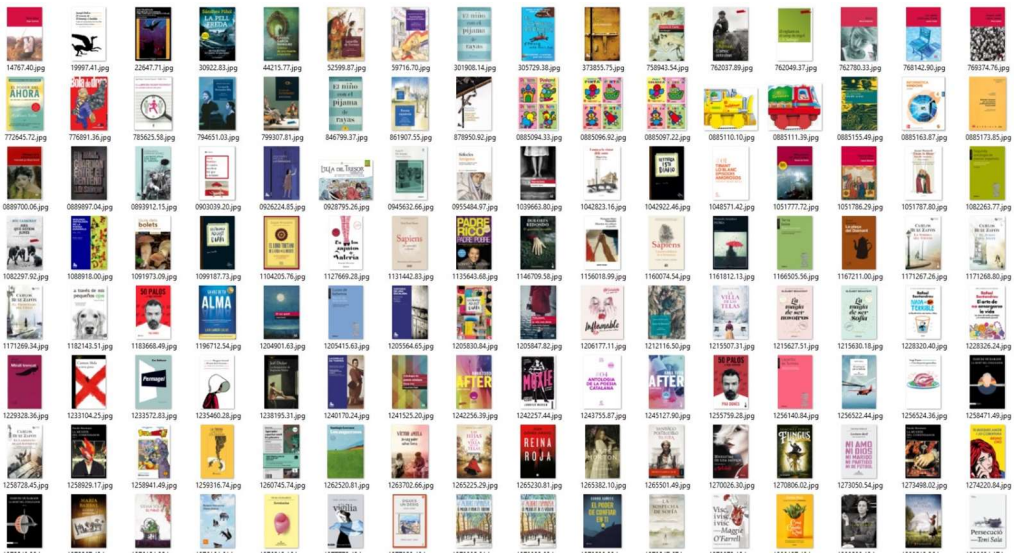
2. ***Definir un titol pel dataset. Triar un titol que sigui descriptiu.***

El dataset extret rebrà el títol de "Catàleg de llibres Abacus Coop."

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

És per això que l'objectiu d'aquesta primera pràctica és extreure la informació disponible sobre tots els llibres que es venen a través de la pàgina web de l'empresa Abacus Cooperativa (https://www.abacus.coop/es/libros), de manera automàtica. Conseqüentment, "La llar del llibre" podrà fer un seguiment constant i s'hi podrà ajustar de manera més precisa amb l'esperança de recuperar-ne les ventes perdudes.

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



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

Les dimensions que s'extrauen sobre els llibres són les següents, en el mateix ordre, guardades en un arxiu CSV (excepte la Imatge, on es guardarà en una carpeta en el directori des de on s'executa el script).

- Títol del llibre
- Autor del llibre
- ID de l'autor: Identificador numèric únic lligat a l'autor
- Editorial: Editorial a la que pertany el llibre
- Preu de soci: Preu venta per als socis de la botiga
- Preu de venta al públic (PVP): Preu venta públic per a qualsevol persona no sòcia.
- Tipus de producte: A banda de llibres, Abacus ven altres productes. En aquest cas ens interessen els llibres.
- Codi EAN: Codi únic lligat al llibre
- Número de referència Abacus: Codi intern d'Abacus que fa referència al llibre
- Enquadernació: Tipus d'acabat del llibre.
- Col·lecció: En el cas que pertany a una col·lecció aquí es fa referència a la mateixa
- Matèria: Aplica en el cas de llibres escolars
- Ruta: Indica la família o tipus al que pertany el llibre. Com a exemples: Novel.la, Còmic, Llibre de Viatge, etc.
- Imatge: Imatge amb la portada del llibre
- Data d'extracció: Cada cop que es realitza una extracció, s'hi afegeix la data de la mateixa.

Tot i que la variable del preu pot semblar la més important per a "La llar del llibre", recomanem col·leccionar també les altres dimensions disponibles, ja que a banda de tenir una influència directe en el preu ens permetran fer posteriors anàlisis.
Aquesta informació extreta serà vigent en el moment en que s'executi el script i és per aquest motiu que s'ha afegit la data d'extracció en el fitxer. En recomanem l'execució periòdica, ja sigui quinzenalment o mensualment, per tal de fer un seguiment de l'evolució de l'oferta editorial i els seus preus.

Per tal d'executar el script en el llenguatge de programació Python, és necessari instal·lar les següents llibreries:

```
pip install pandas
pip install selenium
pip install chromedriver-py
```

Important: és necessari tenir instal·lat el "Chrome Driver" a la mateixa ruta des d'on s'executarà l'arxiu .py.

En l'execució del script es proporcionaran 2 paràmetres:

- Paràmetre 1: L'usuari pot escollir entre les següents classificacions. En cas de no indicar-se, es descarregaran les dades a partir de la pàgina principal (https://www.abacus.coop/es/libros)
    - *top:* Llibres més venuts
    - *novetat*: Llibres en novetats
    - *comic*: Llibres de tipologia còmic
    - *viatge*: Llibres de tipologia viatge
    - *informatica** Llibres de tipologia informàtica
    - *noficcio*: Llibres de tipologia No-ficció
- Paràmetre 2, número de pàgines: Variable numèrica. En cas de no indicar-se o bé indicar 0, es descarregaran les dades de totes les pàgines disponibles.

  Un exemple seria:

  Abacus_Scraping.py top 0

  Com a resultat obtindríem tota la relació de llibres més venuts.

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

Agrair a la Cooperativa Abacus (https://www.abacus.coop/es/home) , propietària de les dades, poder fer ús de les mateixes.

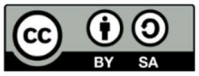
7. Inspiració. Explicar per que es interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Amb aquest tipus d'anàlisis pretenem respondre a les següents preguntes i actuar conseqüentment:

- Evolució de preus (tant PVP com de socis) per llibre i al llarg del temps, la qual cosa permetrà al nostre client adaptar-ne els preus.
- Analitzar la diferència entre Preu PVP i de Soci i considerar la possibilitat d'aplicar un política semblant
- Conèixer quins llibres té l'Abacus Coop. en catàleg i oferir-los si s'escau
- Veure si l'editorial té una influència en el preu de venta
- Veure'n les ofertes puntuals, analitzar cada quan es llancen i quin dia de la setmana per tal d'establir una política d'ofertes dins "La llar del llibre"
- Analitzar les novetats editorials en funció de diversos criteris: preus, editorial, tipus

8. Llicència. Seleccionar una d'aquestes llicencies pel dataset resultant i explicar el motiu de la seva selecció

La llicència escollida ha estat la CC BY-SA 4.0 pels següents motius:



- Es permet fer un ús comercial. Permetria que una empresa emprés les dades generades en projectes que reconeguessin l'autor original.
- Cal mantenir el nom del creador del conjunt de les dades i esmentar els canvis fets respecte la seva versió original.
- Les contribucions es distribuiran segons els paràmetres plantejats pel propi autor.

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

El codi ha estat realitzat amb Python.
Veure arxiu Abacus_Scraping.py

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

El resultat del data set es troba publicat en dues fonts diferents:

- Github: Veure arxiu Abacus_Books.csv
- Zeonodo: 10.5281 / zenodo.4249668
  Font, Carlota y Dil, Jordi. (2020). Llibres Abacus (Versión v2.0) [Conjunto de datos]. Zenodo.
  Veure link https://doi.org/10.5281/zenodo.4249668

Contribucions

Contribucions	Signatura
Recerca Prèvia	Jordi Dil Giró, Carlota Font Castell
Redacció de les respostes	Jordi Dil Giró, Carlota Font Castell
Desenvolupament del codi	Jordi Dil Giró, Carlota Font Castell

Treball realitzat pels estudiants Jordi Dil Giró i Carlota Font Castell