Dataset: Articles de compra del comerç Newluxbrand

David Closa Ortega i Catalin Dediu

Context:

En el món del comerç electrònic un dels factors més importants a l'hora de que un usuari prengui la decisió de compra és el preu del producte. Per aquest motiu, és una pràctica habitual que els comerços monitoritzin els preus dels productes de la competència per a elaborar una estratègia de preus eficient.

A part de la monitorització de preus, també és comú extraure certs atributs de la fitxa de producte de pàgines web de la competència per a enriquir les pròpies fitxes de producte ja que, fent-ho d'aquesta forma, el comerç s'estalvia moltes hores de revisió i redacció de fitxes de producte. Per aquest motiu hem trobat especialment interessant escrapejar tant el preu dels productes com diferents atributs.

Descripció del Dataset:

En aquesta practica s'extreu un conjunt de dades que defineix el producte com el nom del producte, la el preu del producte i la categoria.

A més, també es fa una extracció més detallada del la fitxa tècnica del productes, el color en cas de que sigui disponible i les dimensions.

Representació Gràfica:

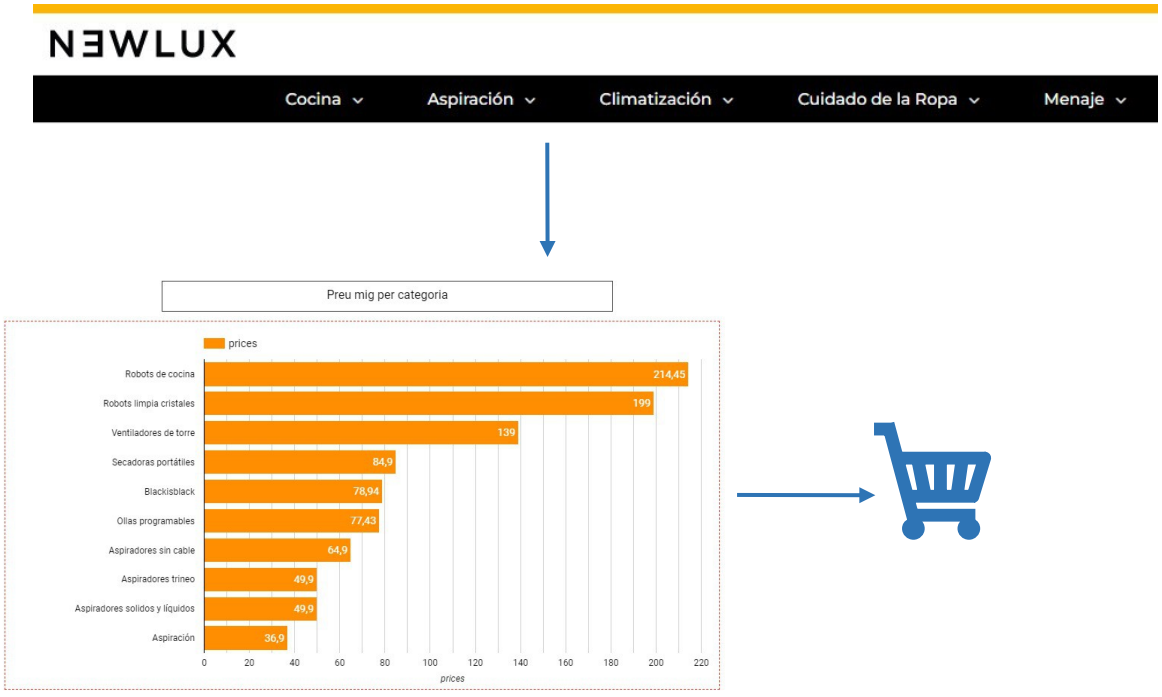

Figura 1: Representació dels 10 productes més cars de la pagina web NewLux

Contingut:

En total s'han extret un total de 6 variables, 1 numèrica i 5 categòriques.

- **title**: nom del producte, variable de referencia del producte
  ex: Robot de cocina Robotmix Multi-Touch RM990

- **princes**: preu del producte en euros, variable numèrica
  ex: 379,0€ (és el preu del producte anterior)

- **categories**: representa la categoria del producte, variable categòrica
  ex: Robots de cocina

- **disponibilitats**: conte la informació tant de si el producte esta en stock com del període d'enviament. Aquesta variable es pot separa en dues per fer un millor tractament de les dades. És una variable categòrica.
  ex: En stock, envio en 24h-48h

- **colors**: representa el color del producte, és una variable categòrica. No tots el productes disposen d'aquesta informació a la fitxa tècnica.
  ex: Negro

- **dimensions**: conte la informació de l'altura, longitud i l'amplada. És una variable categòrica però es pot transforma en 3 variables numèriques. No tots el productes disposen d'aquesta informació a la fitxa tècnica.
  ex: 24x36x37 cm

Agraïments:

Les dades s'han extret directament de la pagina web de l'empresa Newlux, que forma part de l'empresa Mark Join Venture, S.L, amb tècniques de web scrapping.

Inspiració:

Com s'ha comentat prèviament, el conjunt de dades dels productes dels comerços pot ser molt útil tant com pels clients o com per les empreses del sector per realitzar estudis de mercat, comparació de preus, d'estocs etc...

Aquest estudi de mercat pot ser crucial per la rendibilitat de l'empresa ja que pot facilitar les estratègies comercials, l'assignació dels preus amb l'objectiu de ser més competitius. També pot ajudar els comercials a conèixer els detalls el productes ja que no sempre disposen d'aquesta informació.

Llicencia:

La llicencia més adequada per la publicació d'aquest Dataset és la **CC BY-SA 4.0**. El fet determinant de l'elecció d'aquesta llicencia es perquè permet l'ús comercial de les dades. Com que tota l'extracció l'hem enfocat a donar-li un ús comercial considerem que és un aspecte necessari.

A part d'aquesta característica s'ha de tenir en compte que aquesta llicencia permet:

- Compartir, tant com copiar i redistribuir el material
- Adaptar, transformar i crear material amb qualsevol finalitat.
- S'ha de reconèixer l'autoria de manera apropiada mitjançant un enllaç a la llicència i indicar els canvis realitzats.
- En cas de modificacions o transformacions s'ha de publicar amb l'obre resultant s'ha de publicar amb a mateixa llicencia.

Codi:

S'ha utilitzat el llenguatge de programació Python. En Python s'ha utilitzat principalment la llibreria BeatifulSoup per la realització del web scrapping. Aquesta llibreria permet una fàcil negació en el fitxer HTML de la pagina web.

El codi que genera l'extracció és pot trobar en el següent enllaç:
https://github.com/dclosao/tipologia_PRA1

Dataset:

El Dataset es pot trobar a Zenodo a l'enllaç: https://zenodo.org/record/6438650#.YlPFdyhBxD8
amb el codi DOI: 10.5281/zenodo.6438650

Contribucions	Signatura
Investigació prèvia	David Closa Ortega
Redacció de les respostes	Catalin Dediu
Desenvolupament del codi	David Closa Ortega & Catalin Dediu