Màster Universitari en Ciència de Dades

Pràctica 1: Web Scraping

TIPOLOGIA I CICLE DE VIDA DE LES DADES

Antoni Vanrell Valls

Guillem Mir Fuster

30 de març de 2021

1. Context.

L'actual pandèmia provocada per la Covid-19 ha normalitzat l'ús de material higiènic i sanitari de tot tipus a les llars del territori espanyol. Si fa un any la compra d'aquest tipus de material es trobava majoritàriament enfocada a farmacèutiques o negocis especialitzats, avui en dia és una necessitat present al dia a dia de la ciutadania. A més, les restriccions comercials i de mobilitat han intensificat una tendència ja present, les compres online. En aquest aspecte, la pàgina web d'amazon s'ha erigit com a líder indiscutible en el mercat online. És per això que considerem el seu apartat de novetats en salut i cura personal, una font d'informació especialment sensible per conèixer les tendències canviants en aquest mercat.

2. Definir un títol pel dataset.

Últimes novetats en salut i cura personal d'Amazon Espanya.

3. Descripció del dataset.

El dataset generat recull la informació bàsica de les últimes novetats aparegudes a la pàgina de salut i cura personal d'Amazon Espanya. En concret, es recull el nom del producte, les valoracions i el preu, entre d'altres.

4. Representació gràfica.

A continuació anem a veure com estan disposats els elements a la pàgina web.

Si ens fixem amb el primer, observem que l'identificador està a dalt a l'esquerra, una fotografia d'aquest enmig, seguit del nom i la puntuació entre 0 i 5 estrelles. Al costat de la puntuació està el nombre de ressenyes i finalment l'interval de preu.

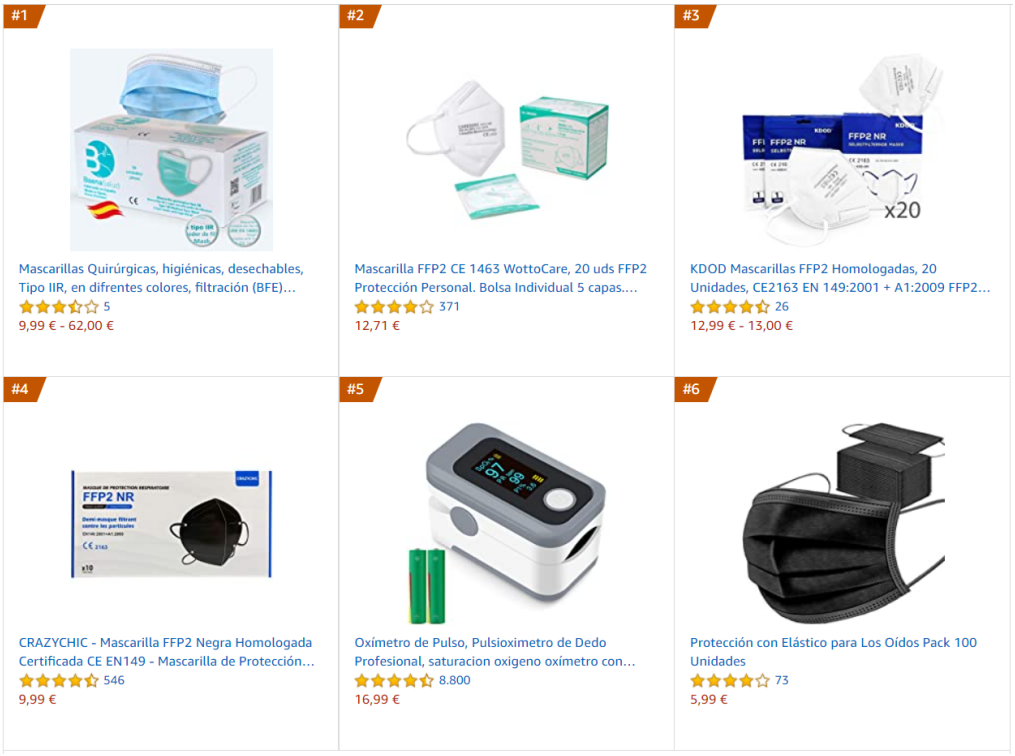

Figura 1: Esquema dels 6 primers elements de la pàgina web d'Amazon: *Últimas novedades en Salud y cuidado personal* a dia 29/03/2021

5. Contingut.

El dataset obtingut consta d'un total de 100 entrades amb 7 camps o columnes:

- **ID:** posició del producte en les darreres novetats en salut i cuidat personal.
- **Nom:** nom complet del producte.
- **Link_Productes:** link del producte per accedir a comprar-lo.
- **Preu:** interval de preu del producte.
- **Puntuació:** puntuació entre el 0 i el 5 mitjana de diferents ressenyes.
- **Ressenyes:** total de ressenyes del producte.
- **Link_Ressenyes:** link amb les ressenyes dels clients.

El dataset recull les últimes novetats en productes a la venta en Amazon en salut i cura personal del dia 29/03/2021.

La pàgina web recull els productes nous amb majors ventes sobre salut i cura personal actualitzant cada hora.

Les dades han estat recollides utilitzant el llenguatge Python per fer Web Scraping per extreure la informació de la pàgina HTML.

6. Agraïments.

Agrair a **Amazon.com** el fet de deixar-nos accedir a la seva pàgina web (`https://www.amazon.es`) per poder extreure la informació que volíem.

7. Inspiració.

Tal com explicat en aquest context de crisi sanitària és especialment important la rellevància que han pres el productes especialitzats en salut en el dia a dia dels consumidors. Aquest conjunt de dades pot servir com una aproximació simplificada a l'oferta generada. A partir d'ell es pot respondre quines són les tendències en l'oferta de productes online de tipus sanitària, així com es valoren i quins preus tenen. D'aquest manera es poden intuir oportunitats d'inversió, conèixer de manera simplificada quina és l'orientació productiva d'aquest mercat o el dinamisme del "branching"en aquest sector.

8. Llicència.

9. Codi.

S'adjunta el codi en python.

10. Dataset.

Pendent de pujar el dataset a Zenodo.

Referències

[1] Laia Subirats Maté i Mireia Calvo González. *Web scraping.* PID_00256968.

Taula de contribucions al treball:

Contribucions	Signa
Recerca prèvia	GMF, AVV
Redacció de les respostes	GMF, AVV
Desenvolupament codi	GMF, AVV