Elaboració d'un dataset amb productes de la web electrocosto.com mitjançant web scraping

1. Context

Avui en dia les compres per internet són gairebé una cosa rutinària. Els clients poden accedir de forma fàcil i ràpida a una extensa quantitat de botigues en línia. Totes aquestes facilitats juntament amb campanyes de màrqueting molt agressives afavoreixen les compres impulsives. On un client es pot veure comprant un producte que una estona abans no sabia ni que existia ni que el volia. Aquest tipus de conductes no es produeixen amb tota classe de productes, ni de preus. El més habitual és que els usuaris pensin i investiguin més quan es tracta de productes d'un preu més elevat, és comú llavors llegir ressenyes d'altres clients i mirar vídeos de webs o canals especialitzats, per assegurar-se que el producte satisfarà totes les necessitats i conèixer per avançat que és el que podem esperar d'ell. És en aquest context, quan les persones solen dedicar més temps a comparar els preus de diferents botigues per tal de comprar el producte que han escollit al menor preu possible.

Si observem la situació des de l'altra banda, és a dir des del costat d'una botiga observarem que oferir els preus més baixos del mercat pot aportar moltes vendes. Per altra banda, tampoc cal oferir un gran descompte respecte a la resta de botigues que venen el mateix tipus de productes, ja que estariem reduint els nostres beneficis. Sota aquestes condicions, es dóna el cas que articles de molta tirada es troben al mateix preu a moltes botigues ja sigui perquè el preu ve marcat pel fabricant o perquè les botigues volen oferir preus competitius alhora que maximitzen beneficis.

No és el cas d'altres productes de menys tirada on podem trobar diferències significants de preus entre les diverses botigues. Sobre aquest tipus de productes proliferen l'aparició de webs que es dediquen a comparar el preu dels diferents productes a les diferents webs per tal d'ajudar al comprador a trobar els millors preus. Tot i que els comparadors de preus poden ser molt útils per als consumidors a vegades poden ser insuficients per una botiga que vol estudiar els seus competidors. Tant des del punt de vista d'un comparador com d'una botiga pot resultar molt útil rastrejar els productes i els preus de les botigues en un determinat sector així com les característiques de productes. Recollir aquesta informació permet a les botigues posar preus competitius a productes amb característiques semblants, per part del

recomanador permet la creació de buscador sofisticat on es puguin introduir paràmetres per acotar la cerca així com recomanar productes semblants que es troben a millors preus.

Un dels sectors que reuneix les característiques anteriorment descrites és el dels electrodomèstics. Tothom necessita comprar electrodomèstics de tant en tant, es tractar d'una compra premeditada i que sol venir acompanyada d'una recerca a l'alçada del preu.

Després de realitzar una recerca exhaustiva sobre botigues en línia d'electrodomèstics vam trobar la web https://www.electrocosto.com/ una botiga especialitzada en electrodomèstics i que només realitza comandes en línia i envia a domicili l'electrodomèstic comprat. Disposa d'una àmplia gamma d'electrodomèstics, és per això que vam decidir triar aquesta web perquè en un mateix lloc podem trobar informació de neveres, televisors, microones, cuines, rentadores, etc. Aquesta web ens permet dur a terme un cas pràctic del rastreig d'una botiga que hem comentat fa uns instants. La informació que podem extreure d'aquesta web pot ser d'un alt valor per a recomanadors o altres botigues del sector. Un projecte més ambiciós podria valorar els avantatges de rastrejar diferents webs que es dediquin a la venda del mateix tipus de productes i recopilar la informació provinent de les diferents fonts

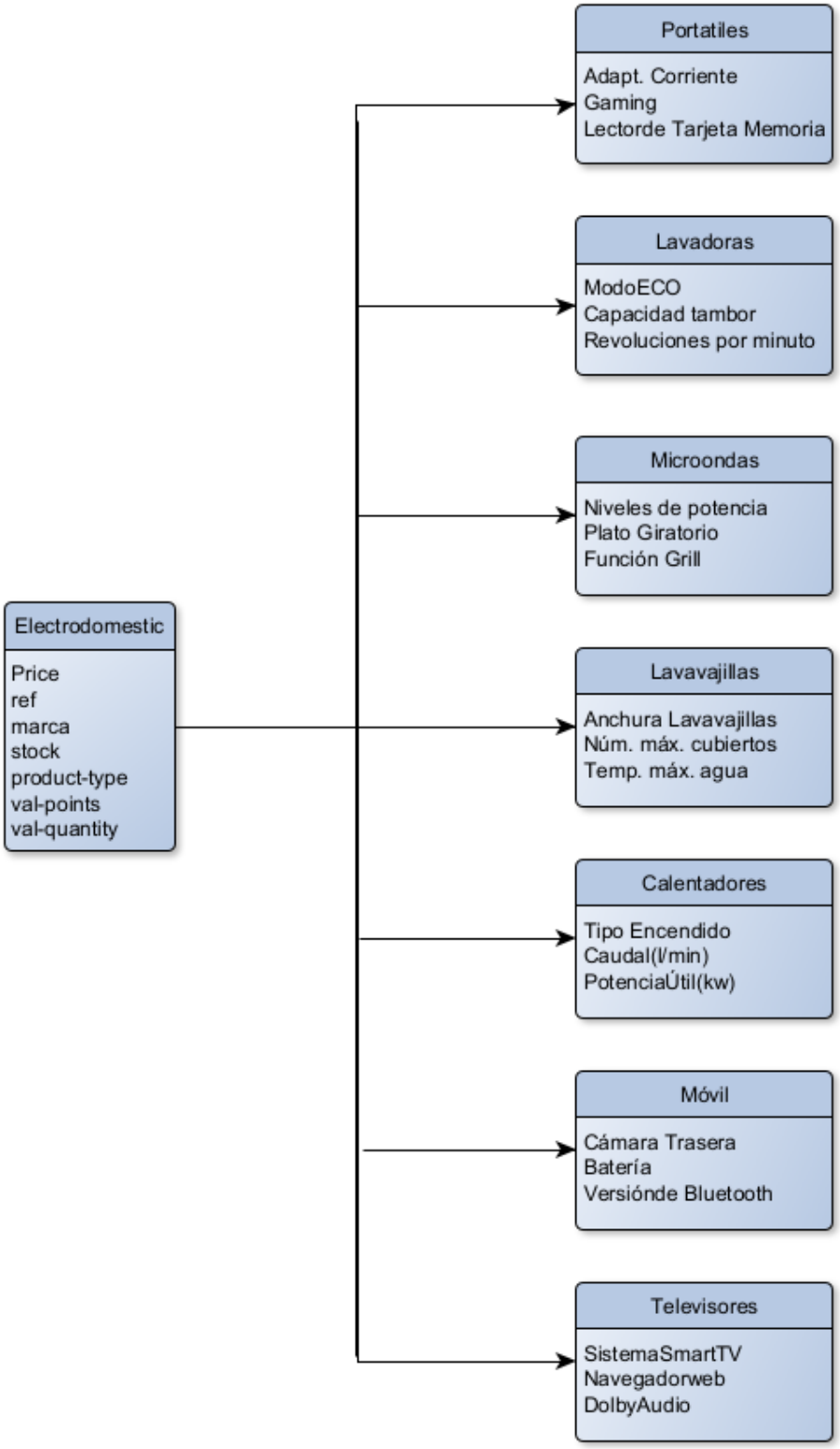
2. Títol

Productes i característiques del catàleg d'electrodomèstics de la web electrocosto.com.

3. Descripció del dataset.

El dataset consta de les diferents característiques, preus i informació tècnica del conjunt de neveres ofertades a la pàgina web de venta d'electrodomèstics electrocosto.com.

4. Representació gràfica



5. Contingut

El dataset té dues parts, una part comuna i una part específica. A la part comuna tenim camps que apliquen a tots els tipus de dispositius que poden haver-hi al dataset. Tenim:

- product-name: Nom del producte.
- product-type: Tipus d'electrodomèstic (segons la classificació pròpia de la web)
- price: Preu en el moment de l'extracció de dades.
- ref: Referència del producte.
- marca: Marca/empresa que fabrica el producte.
- val-points: Valoració 1-5 de la qualitat del producte segons els usuaris.
- val-quantity: Nombre de valoracions.
- stock: Si el producte està o no disponible.
- product-sending-value: Costos d'enviament.
- date: Data de l'extracció de dades.

A la part específica tenim totes les característiques del producte que apareixen a la web. Com no tots els productes les tenen fins i tot sent productes del mateix tipus disposen de molts camps nuls i n'hi ha moltes de diferents. No les detallarem totes però alguns exemples serien:

- ModoECO: Té mode ECO o d'estalvi.
- CapacidadTotalNeta: Capacitat de volum en Litres.
- Cajones: Disposa de calaixos.
- Peso: Pes.

Les dades pertanyen al període de temps especificat en el camp "date". D'aquesta manera podem tenir registre de l'evolució dels productes al llarg del temps en aquesta web.

6. Propietari

El conjunt de dades és propietat d'Electrocosto, la web d'on les hem extret. Tal i com s'explica a Henrys, K. (2021). Importance of web scraping in e-commerce and e-marketing. *Available at SSRN 3769593*., avui dia disposem d'una quantitat enorme de dades a les xarxes de les quals amb una mica de dedicació, podem extreure molta informació d'elles. Altres estudis similars s'han realitzat per exemple amb els lloguers

de cases i apartaments als estats units: Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. *Journal of planning education and research*, *37*(4), 457-476. i amb el preu dels pisos al regne unit: Bricongne, J. C., Meunier, B., & Pouget, S. (2023). Web-scraping housing prices in real-time: The Covid-19 crisis in the UK. *Journal of Housing Economics*, *59*, 101906.

7. Inspiració

Com hem mencionat anteriorment, el dataset consta de les característiques de diferents productes electrodomèstics d'una web de compra, concretament ens hem centrat en les neveres però amb fàcil capacitat d'escala.
El dataset es pot enfocar de dues maneres: estudiar l'evolució al llarg del temps d'aquests productes, si varia el preu, s'afegeixen de nous, l'estoc està disponible o no, valoracions dels usuaris etc, el segon punt de vista pot ser com a comparació entre els propis productes, quin té millors valoracions, ordenació per preu, més característiques, marques que tenen més models diferents etc.
D'aquesta manera podem filtrar i extreure informació que requeriria d'una feina adicional als usuaris de la web de cara a ajudar a decidir quin producte necessiten, o podem extreure comparatives amb altres estudis que es puguin fer més endavant amb altres webs per per exemple decidir quina pàgina ofereix els millors preus.
De cara a la part ètica hem decidit afegir un delay entre les diferents requests que fem a la pàgina web per evitar saturar els servidors. També comprovem que no accedim a cap url que no estigui permesa per l'arxiu robots.txt.

8. Llicència

9. Codi

Com totes, la nostra pàgina web està dissenyada per la interacció amb les persones, el que vol dir que ens hem hagut d'adaptar als diferents mecanismes/scripts que se'ns presenten.
Primer de tot li hem de proporcionar al script a quina secció d'electrodomèstics volem accedir. D'aquesta manera no hem de fer cap navegació inicial i ens porta a una pàgina on apareixen objectes d'aquest tipus:

Aquests blocs ens donen una informació resumida del producte i contenen l'enllaç de la pàgina pròpia del producte. El nostre objectiu és recopilar el total de les url de tots els productes per després poder procedir a l'extracció de les dades. Hi ha un problema, però, a la pàgina inicialment només es mostren els primers 16 productes, la resta queden amagats i només apareixen en fer scroll cap a baix i després d'un parell d'scrolls et demana que premis un botó per acabar de desplegar els productes que falten.

Hem hagut de treballar al voltant d'aquesta interacció amb la llibreria Selenium, que ens permet obtenir l'HTML de la pàgina després d'haver desplegat tots els productes per poder tot seguit extreure amb un script tots els links d'aquests. Els scrips relacionats amb aquest tipus d'interacció i que gestionen el contingut que es genera de manera dinàmica els podem trobar a l'arxiu de Python seleniumScript.py, aquest scrip pot no funcionar si l'executable de Firefox no es troba en el lloc habitual. La principal funció

d'aquest arxiu s'anomena expand_section, donada el nom d'una secció de la web retorna l'HTML de la web amb tots els electrodomèstics d'aquella categoria desplegats i llest per extreure els links que porten a cada un dels ítems. La resta de funcions es tracten de funcions auxiliar a aquesta.

Finalment, arribem a la part de l'extracció de les dades. Com hem mencionat a l'apartat del dataset, tenim dues parts, una general i una específica:
La part general representa l'estructura que tenen tots els productes per igual, simplement examinant ràpidament el codi HTML podem extreure sense cap problema les diferents característiques de cada producte.

La part específica és més complexa, a la web es representa com una taula amb el nom de la característica i el camp amb la informació. El problema és que no tots els productes tenen el mateix nombre de camps. Per evitar doncs problemes amb les dimensions del dataset, un cop tenim la llista de links, fem una lectura prèvia per establir l'estructura del dataset afegint tots els camps possibles de tots els productes. A continuació afegim els productes i anem comprovant que els camps on anem afegint informació continguin el nombre d'elements que toca i afegint nuls en cas que no sigui així. Tota aquesta informació que extraiem es guarda en forma de diccionari que al final convertim en format CSV. Aquest procés es realitza a l'arxiu de Python electrocosto_scraper.py. La funció product_scrapper s'encarrega de gestionar les sol·licituds a les diferents seccions, així com de crear el diccionari i d'extraure les característiques dels electrodomèstics i incorpora-les de forma adequada al diccionari. La funció download s'utilitza per descarregar el codi HTML de cada una de les webs

dels diferents productes. La resta de funcions serveixen de funcions auxiliars a aquestes.

CARACTERÍSTICAS GENERALES	
Cámara	No
Enfriamiento rápido	Sí
Motores	1 Motores
NoFrost	Total
Número de puertas	2 Puertas
Panelable	No
Refrigeración	No Frost
Modo Vacaciones	No
Tipo de instalación	Libre
Modo ECO	Sí
Motor Extrasilencioso	Sí
Tipo de Botellero	Rejilla auxiliar para 1 botella
Capacidad de congelación	10 Kg/24H

10. Dataset

Podem trobar el dataset en l'apartat /dataset al repositori de GitHub https://github.com/afinolUOC/electrodomestics-scraping o de formar semblant es pot descarregar el dataset a Zenodo a través del següent enllaç del DOI: https://doi.org/10.5281/zenodo.7863333

Hem considerat oportú pujar dos datasets:

- neveras.csv: Conte la informació sobre totes les neveres que podem trobar a la pàgina web electrocosto.com. Permet l'elaboració d'un model de predicció de preus degut a que tots els electrodomèstics pertanyen a la mateixa categoria.
- electrodomestics.csv: Conte l'informació sobre tots els electrodomestics que podem trobar en les seguents categories: lavadoras, microondas, lavavajillas, moviles, televisores, portatiles i calentadores. Permet l'elaboració de models de classificació d'electrodomèstics o realitzar agrupacions diferents de les creades per les seccions d'electrodomèstics.

11. Video

Contribucions	Signatura
Investigació prèvia	AFP, AGV
Redacció de les respostes	AFP, AGV
Desenvolupament del codi	AFP, AGV
Participació al vídeo	AFP, AGV