Ernest Panareda Roig
Claudia Puche Garcia

PRÀCTICA 1 - TIPOLOGIA I CICLE DE VIDA DE LES DADES

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

En un context de constant canvi i grans quantitats d'informació generades cada segon trobem que és de vital importància tenir un accés a la informació viable. És per aquesta raó que la tècnica de Web Scraping és molt útil quan es tracta de recol·lectar dades de qualsevol pàgina web. Aquest mètode va sorgir per la necessitat de tenir les dades d'un lloc web d'una manera ràpida i fàcil, a més de tenir-les constantment actualitzades. Per aquesta pràctica hem aplicat aquest procediment al lloc web d'Amazon, ja que és una companyia que ofereix gran diversitat de productes i que opera al voltant del món. Concretament, al 2020 es comptabilitzen 166 milions de productes oferts de 20 categories diferents; a més, mostra la informació de quantes persones han valorat el producte i amb quina puntuació. Allò que ens ha cridat l'atenció d'Amazon és la llista coneguda com Best Seller Rank que agrupa 100 productes que han experimentat una tendència a l'alça de vendes contemplant el seu històric continuat en el temps. Per tant, hem elegit aplicar el web scraping a la llista de Best Seller Rank d'Amazon, ja que ens dona informació constantment actualitzada de quins són els productes més venuts cada hora de cada categoria.

Aquest és un mètode útil per descobrir productes amb èxit. Quan s'investiguen productes per vendre a Amazon, s'ha de tenir en compte la seva demanda i la llista dels més venuts pot proporcionar aquesta informació. De la mateixa manera, també pot ser realment eficient si l'objectiu és aconseguir informació de la competència. El rànquing pot ajudar facilitant el seguiment dels principals competidors mirant productes similars. Mitjançant les dades extretes es podrien reajustar mesures com la modificació de paraules clau o la dinamització dels productes venuts. Finalment, la funció més utilitzada del Web Scaping aplicat a la BSR és la predicció de les vendes. Per exemple, si venem un producte que es troba a la posició 2000 de la llista i equival a unes 700 vendes al mes es pot determinar que un producte llistat a una posició superior tindrà una demanda superior i unes vendes més elevades. Tot i així, s'ha d'anar amb precaució perquè la classificació pot variar molt entre categoria. Per aquesta raó, hem realitzar web scraping de totes les categories d'Amazon.

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

Els més venuts d'Amazon (*Amazon best sellers*, utilitzat en la publicació, ja que s'ha fet en anglès).

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

Com s'ha comentat anteriorment el dataset extret recull informació sobre la classificació de Best Seller Rank. En concret, les variables que formen el dataset són 7 i constitueixen les variables que normalment influeixen en el criteri de compra de cada consumidor. Aquestes variables són data, categoria, rànquing, producte, estrelles, opinions i preu. Per exemple, tot comprador d'Amazon ha tingut en compte les estrelles del producte o els comentaris que han postejat els anteriors consumidors. Per tant, les variables seleccionades ens poden donar una visualització de com funciona aquest rànquing per categoria a Amazon. Addicionalment, una de les variables també és la data en la que es recull la informació, que primerament no sembla gaire rellevant, però que si tenim en compte que la classificació es va actualitzant cada hora i que pot ser el que interessa és l'evolució d'un cert producte és una constant important.

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

Hem seleccionat aquesta imatge, ja que representa gràficament el Best Seller Rank d'Amazon. A més, mostra com gràcies a la informació els venedors poden augmentar la seva posició a la classificació.

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

Les dades compten amb set variables explicades a continuació:

- Data: data en la que es va recollir les dades d'aquell producte
- Categoria: categoria de cada producte, per exemple Alimentación y bebidas
- Ranking: posició que pren el producte al rànking d'Amazon
- Producte: nom identificatiu del producte
- Estrelles: ponderació del 0 al 5 de les valoracions dels consumidors del producte
- Opinions: número d'opinions i comentaris registrats
- Preu: preu del producte en euros

Les dades daten un 28 d'octubre que va ser el dia que es va executar l'script tot i que com s'ha mencionat anteriorment la idea és utilitzar aquest codi per realitzar una automatització i extreure dades en un moment determinat del temps.

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

Les dades van ser extretes de la pàgina web de Best Seller Rank d'Amazon mitjançant codis de programació en llenguatge python. A través de la llibreria de Beautiful Soup i de requests vam emprar tècniques de Web Scraping per extreure la informació necessària per la PRA 1 de l'assignatura.

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

Aquest conjunt de dades és interessant ja que permet conèixer la situació del mercat en línia sobre els productes més demandats en un moment determinat en el temps. Amazon és l'empresa número 1 de vendes en línia, que a més no té un únic proveïdor, un únic tipus de producte, sinó que ofereix de tot i de tothom qui en contracti el seu servei. Així doncs, els seus indicadors no són només útils per la venta de productes en línia sobre aquesta plataforma, sinó que es poden extrapolar a la venta en línia en general i a la venta física en establiments.

Fet un resum de la definició de què és Amazon, es pot fer una idea del potencial que tenen les dades dels productes que estan en els 100 més venuts.

El dataset que s'obté de fer scraping sobre la web d'Amazon presentat en aquesta pràctica, conté els 100 productes més venuts el dia 28 d'octubre de 2020 sobre 35 categories diferents, des de menjar i beguda o llar i cuina, fins a informàtica o música. Així doncs, d'aquest dataset es pot extreure informació útil sobre diferents sectors per saber les tendències del moment i poder enfocar estratègies de negoci.

L'aplicació més fàcil d'aquest dataset és per una proveïdor que vulgui vendre un producte per Amazon, aleshores, en base als productes de que disposi i la informació de les dades, aquest serà capaç de saber si contractar el servei d'Amazon com a punt de venta del seu producte serà rentable o no, buscant per al mateix producte en la llista i/o productes similars, la posició en el top 100, el preu al que es venen i les valoracions dels compradors.

Però a nivell més genèric, una empresa que ofereix un producte o té previst llançar-ne un de nou, sigui del sector que sigui, pot extreure informació dels productes que més s'estan venent en el moment actual i compara per saber si aquesta oferta és competent al mercat, o en el cas del llançament d'un de nou, adaptar-lo a la demanda actual, o potser fer un replantejament de si seguir amb aquest projecte del llançament del producte perquè les dades no són favorables, i així acabar definint una estratègia lògica i d'èxit, fonamentada en la informació dels productes més venuts de la empresa en ventes en línia número 1 al món.

En definitiva, les preguntes a que es pretén donar resposta amb aquest dataset són: **És rentable vendre un producte determinat per Amazon? Quins productes són tendència en aquest moment i m'asseguren l'èxit adoptant-los com a negoci? La meva oferta és competent en el mercat actual?**

Finalment afegir que el dataset que s'ha publicat per aquesta pràctica es correspon a dades d'unicament 1 dia, així que per donar resposta a les preguntes que es plantegen aviat es quedaran obsolets. Tot i això, el codi presentat genera automàticament un nou dataset amb les dades de les mateixes però del dia en que s'ha llançat, podent tenir la informació sempre actualitzada al moment.

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

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**
- **Unknown License**

Aquest punt ha estat força delicat, ja que la legalitat que envolta el web scraping presenta el que es "pot fer o no" d'una manera molt ambigua.

Com a primer punt s'ha estudiat el 'robots.txt' (https://www.amazon.es/robots.txt), on s'hi llisten aproximadament un centenar de directoris on no es desitja que s'hi realitzi web scraping, i únicament 7 que sí. Per sort, en aquest document no hi apareix el directori que s'utilitza per fer el scraping, sent aquest '/gp/bestsellers/'.

Per altra banda, en l'apartat de 'Condiciones de Uso y Venta' de 'amazon.es' (https://www.amazon.es/gp/help/customer/display.html?nodeId=GLSBYFE9MGKKQXXM) apareix en el punt 6 'Licencia y acceso' el següent text: "*Sujeto a tu cumplimiento de estas Condiciones de Uso y las Condiciones Generales de los Servicios aplicables, [...] te conceden una licencia limitada no exclusiva, no transferible y no sublicenciable, de acceso y utilización, a los Servicios de Amazon para fines personales no comerciales. Dicha licencia no incluye derecho alguno de reventa ni de uso comercial de los Servicios de Amazon ni de sus contenidos,* ***ni derecho alguno a compilar ni utilizar lista alguna de productos, descripciones o precios.*** *Tampoco incluye el derecho a realizar ningún uso derivado de los Servicios de Amazon ni de sus contenidos, ni a descargar o copiar información de cuenta alguna para el beneficio de otra empresa,* ***ni el uso de herramientas o robots de búsqueda y extracción de datos o similar***". Així doncs, a la web s'exposa que acceptant les seves condicions d'ús no tens dret ni a fer scraping ni a llistar noms, descripcions i preus de productes, sent tot això el que es realitza en aquesta pràctica. Per sort, com s'indica a la teoria, en cap moment s'accepten condicions, així que legalment, aquest cas quedaria exempt. A més, la finalitat d'aquest scraping és purament acadèmic.

Aleshores, prenent la teoria, el nostre cas de web scraping sobre Amazon compleix el protocol d'exclusió de robots, no es salta termes ni condicions (ja que no se n'accepta cap), per tant, únicament rastreja informació pública, no causa d'any al ser una consulta a petita escala, i l'ús de les dades que s'extreuen és únicament acadèmic.

Fins aquí s'ha plantejat la legalitat del scraping, però no de les dades. Investigant el lloc web d'Amazon, d'on s'obtenen aquestes dades, no s'ha trobat cap informació respecte a la llicència de les seves dades de productes. Segurament es deu al fet de que es basa en el marc legal de que extreure dades d'aquest lloc està prohibit.

Paral·lelament s'ha observat que hi ha diferents debats (sobre Amazon i un lloc web en general) sobre si publicar dades obtingudes fent scraping es surt de la legalitat o no, i sota quina llicència s'haurien de publicar, i en aquests debats es mostren diferents opinions al respecte, però predomina l'opinió de que és una acció il·legal.

Arribats a aquest punt, en base a la teoria sobre la legalitat de les dades descrita, que aparentment es compleix en tot moment, la idea seria seleccionar una llicència 'CC0: Public

Domain'. Però com que aquest cas, tot i la teoria exposada, és un cas pràctic i no es té permís per escrit del propietari de la pàgina, **s'ha optat per prendre una decisió conservadora seleccionant la llicència 'Unknown License'**, explicant el marc en que s'han obtingut i publicat les dades en la descripció del dataset, remarcant que és purament acadèmic. Tot i que sota aquesta llicència les dades no son 'Open Data', d'aquesta manera s'evita qualsevol possible repercussió legal que pugui sorgir per la publicació d'aquestes.

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

El codi utilitzat es troba adjunt en el repositori obert de github.

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

El DOI del dataset publicat a Zenodo és el **10.5281/zenodo.4256634.**

CONTRIBUCIONS	SIGNA
Recerca prèvia	CPG, EPR
Redacció de les respostes	CPG, EPR
Desenvolupament del codi	CPG, EPR