Scraping de pàgines de bellesa i cosmètica natural

Context

Davant de la importància que estan agafant els productes vegans i naturals en un context mundial, on la gent està agafant consciència del maltractament animal i del perjudici que això causa al nostre entorn, i en conseqüència, al nostre món, hem decidit realitzar un scraping d'aquesta tipologia d'empreses. Entenem que aquesta tècnica, el web scraping, serveix per a realitzar un bon màrqueting i donar visibilitat i comparacions a les millors ofertes que aquestes empreses ofereixen; donar visibilitat i donar consciència a la població que aquests productes poden ser igual d'eficaços que els productes que contenen químics afegits i/o estan testats en animals.

Descripció del dataset

Els datasets extrets són de dues pàgines de cosmètica i bellesa natural:

- https://www.freshlycosmetics.com/es/productos/
- https://kriim.com/collections/all_collections

Analitzant aquestes pàgines web ens hem adonat de que podem extreure dades, tant per realitzar comparacions entre empreses, com per realitzar anàlisis estadístics d'aquestes empreses individualment.

El dataset conté els noms i aplicacions de diferents productes de cosmètica d'origen natural. Tanmateix, s'hi inclouen els preus, les valoracions i el número d'opinions que té cada producte.

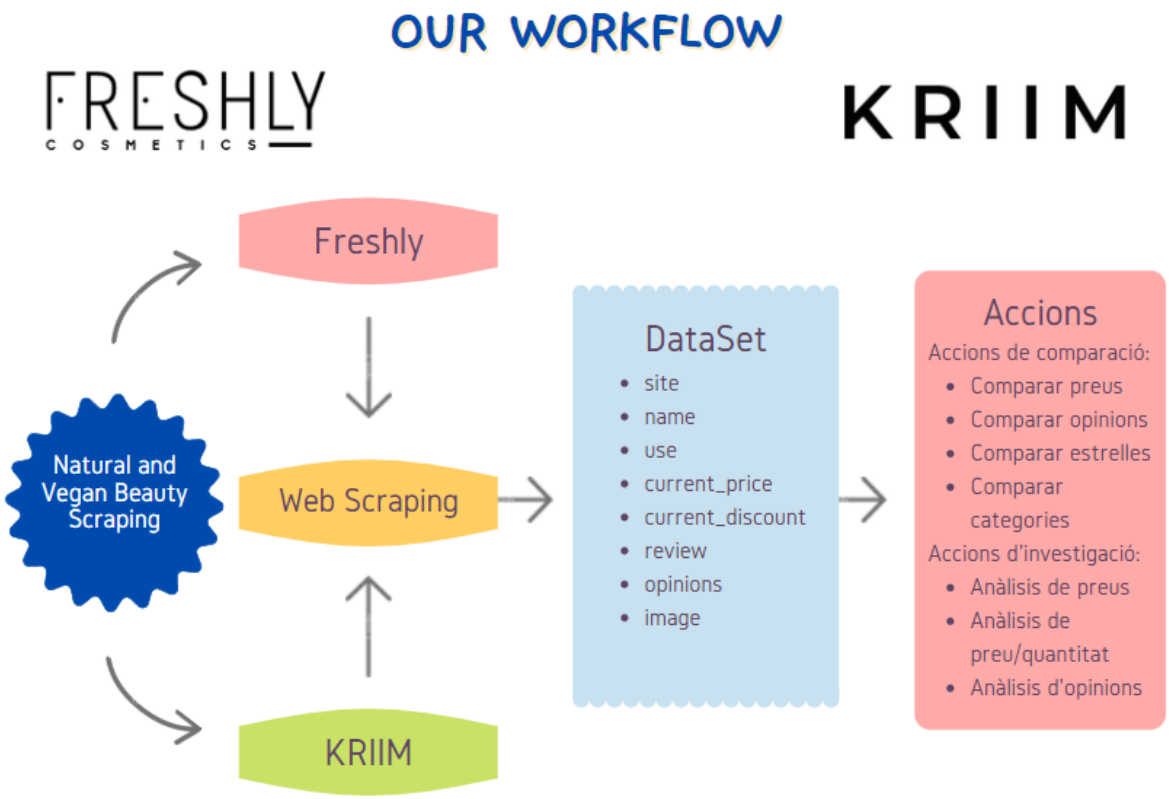
La proliferació de les plataformes de comerç electrònic fa que avui dia, com a consumidors, tenim a l'abast una gran quantitat de plataformes on trobar els mateixos tipus de productes. És per això que, per poder realitzar una compra informada, és indispensable comptar amb el binomi preu-opinió.

El fet d'afegir també un camp referent a l'ús de producte permet la categorització i agrupació de les dades en diferents grups, de manera que es pot establir un nivell de granularitat en el tipus d'anàlisi que es pot arribar a fer amb aquestes dades.
Les dades recollides en aquest dataset són una fotografia del moment en què es van extreure. Aquest projecte no s'ha considerat com un projecte d'scraping recurrent. Tot i això, s'ha inclòs un camp amb el possible descompte aplicat sobre els productes a la venda en els dos portals estudiats.

Títol del dataset

Per els motius anteriorment esmentats, el títol d'aquest dataset és "***Product prices in natural cosmetics ecommerce webpages***".

Representació gràfica



Contingut

Aquests datasets contenen les mateixes variables, que són les següents:

- ***site:*** etiqueta que conté el lloc de descàrrega de la informació del producte
- ***name****:* nom del producte
- ***use****:* ús per al que està pensat el producte
- ***current_price****:* preu actual
- ***current_discount****:* descompte que en aquest moment té el producte
- ***review***: conté el rating mig en estrelles del producte, entès com un valor numèric de 1 a 5.
- ***opinions:*** nombre de opinions d'usuaris que han valorat l'article.
- ***image:*** imatge del producte en format array

Aquesta selecció de variables està pensada per realitzar un futur anàlisis estadístic, com també per poder comparar i fer *joins* amb altres datasets d'altres pàgines de cosmètica.

	site	name	use	current_price	current_discount	review	opinions
0	kriim	Bamboo Hand Treatment	Tratamiento de bambú para manos y uñas	€14,50	9.76	5.0	6
1	kriim	Loss of Elasticity Skin Renewer	Crema antienvejecimiento para piel seca y sensible	€56,95	45.56	5.0	1
2	kriim	Age Delay Eye Concentrate	Concentrado AGE-DELAY para el contorno de ojos	€53,95	53.95	5.0	1
3	kriim	Carota Sativa Cleanser	Limpiador con carota sátiva ideal para pieles sensibles	€9,00	9.0	3.6666666666666665	3
4	kriim	Anti-Acne Serum	Sérum Anti-Acné	€35,00	23.8	4.842105263157895	38
5	kriim	Anti Dark Circles Peptide 4% + Caffeine	Sérum reductor de ojeras y bolsas	€29,50	29.5	4.448275862068965	58

Imatge 1: Dataset de la pàgina web de Kriim

	site	name	use	current_price	current_discount	review	opinions	image
0	freshly	Golden Radiance Body Oil	Aceite corporal 100% natural nutre, reafirma y trata estrías	26,00 €		4.8	7814 opiniones	[[[255 255 255] [255 2
1	freshly	Glow Edition Body Oil	Iluminador corporal natural con destellos dorados	29,00 €		4.5	1405 opiniones	[[[255 255 255] [255 2
2	freshly	Azelaic Radiance Face Treatment	Tratamiento facial imperfecciones, acné y piel reactiva	35,00 €		4.6	3559 opiniones	[[[255 255 255] [255 2
3	freshly	Bronzing Radiance Self-Tanning Cream	Crema autobronceadora, potenciadora del bronceado	29,00 €		4	1712 opiniones	[[[255 255 255] [255 2
4	freshly	Hair Growth & Density Treatment	Sérum de crecimiento capilar	29,00 €		3.9	619 opiniones	[[[255 255 255] [255 2
5	freshly	Hair Radiance Keratin Spray	Spray capilar protector y reparador para un cabello sano, brillante y sedoso	16,00 €		4.2	2802 opiniones	[[[255 255 255] [255 2

Imatge 2: Dataset de la pàgina web de Freshly

Agraïments

Les dades d'aquest projecte són propietat de FRESHLY COSMETICS i de KRIIM. Tot el projecte es basa en una única petició puntual al catàleg públic de la pàgina web, únicament amb finalitats acadèmiques per a la pràctica del web scraping en l'àmbit exclusiu d'aquesta entrega.

En la mesura del possible, s'ha tractat d'anonimitzar les dades obtingudes des del portal KRIIM, ja que presenta la política més restrictiva dels dos llocs web d'estudi. Per fer-ho, s'han randomitzat els valors de current_price, current_discount i review. Tanmateix, no es fan disponibles les imatges.

No obstant això, tota la informació extreta està disponible de cara al públic en el catàleg dels productes, sense necessitat de fer login dins de la plataforma, ni l'acceptació obligatòria de termes i condicions, més enllà de la pròpia visita al web.

Per tal de minimitzar l'impacte de la petició de descàrrega, s'ha establert una espera de 3 segons en la descàrrega de les dades entre productes.

Com s'ha comentat en punts anteriors d'aquest document, la proliferació de l'ecommerce fa que els consumidors es tornin més exigents a l'hora de comprar productes, a causa de la gran quantitat de plataformes i informació disponible.

Tanmateix, a l'hora de posicionar-se en el mercat, el web scraping pot permetre també l'elaboració de l'estudi de mercats i posicionament a l'hora de llançar una nova línia de productes.

És molt habitual trobar comparadors de preus, ja sigui en plataformes tipus "Rastreator" o similar, o bé com a comparacions elaborades per usuaris individuals en pàgines de màrketing d'afiliació.

Cal remarcar la existència d'algunes empreses privades, que profesionalitzen el web scraping com a producte de valor.

Finalment, com agraïment pur, volem agrair al YouTuber John Watson Rooney, aquest YouTuber té una gran quantitat de vídeos ensenyant *web scraping*, amb una gran quantitat de variants per realitzar-ho.

Inspiració

Per tal de realitzar un bon màrqueting, i/o un bon posicionament, entenem que les dades a tractar han de ser, primer, dades comparables, és a dir, categories dels productes. També és molt interessant el preu dels productes per posteriors anàlisis estadístics, com també és interessant les opinions i quantitat d'opinions que el producte ha generat.

El dataset presentat en aquesta pràctica, al cap i a la fi, és un dataset limitat, però amb exposició a diferents tipus de productes, categoritzats segons el seu ús, per tal de permetre establir un punt de comparació.

A més, els productes cosmètics estudiats en aquest dataset solen tenir noms bastant explicatius en quant als seus components principals, de manera que també és informació útil per a l'usuari que realitzi l'estudi.

Les preguntes principals que es poden respondre amb aquest tipus de datasets serien les següents:

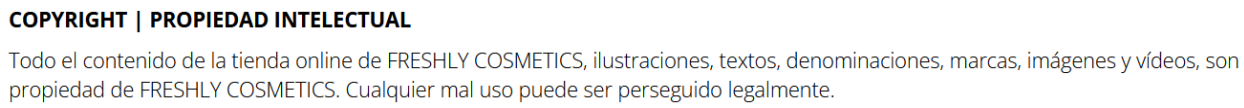
- Quines plataformes posen a la venda productes al millor preu?
- Està relacionat el preu del producte amb la quantitat i puntuació de les opinions dels usuaris?
- A quins usos estan destinats els productes amb major preu? I els de menor preu?
- Quins productes són els més populars entre els compradors?

Com hem comentat en el punt anterior, aquest tipus de preguntes estan orientades també al tipus de posicionament que es vol fer, i per això no és estranya la proliferació de plataformes de comparació entre diferents productes i serveis. En el cas de productes de retail, les comparacions poden ser més senzilles, però aquest enfoc es pot també tenir per a l'estudi de

serveis més complexos. Per exemple, un projecte similar pot ser l'estudi de preus, condicions i prestacions de productes d'assegurances, lloguer de vehicles, etc.

Llicència

La llicència de les dades és *Unknown License*, però realitzant una búsqueda de les condicions d'ús ens fixem en el següent:



COPYRIGHT | PROPIEDAD INTELECTUAL

Todo el contenido de la tienda online de FRESHLY COSMETICS, ilustraciones, textos, denominaciones, marcas, imágenes y vídeos, son propiedad de FRESHLY COSMETICS. Cualquier mal uso puede ser perseguido legalmente.

Imatge 3: Copyright de Freshly

Sota les condicions més estrictes, el llicenciament d'aquestes dades anonimitzades només podrien considerar-se sota la llicència Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). Tot i això, el més adient seria un llicenciament desconegut, sempre tenint en compte les següents consideracions:

Aquest és un projecte purament acadèmic, on s'estudia com realitzar la obtenció de dades des de pàgines web mitjançant tècniques de web scraping. Tots els usos de les dades queden englobats únicament en la finalitat pròpiament personal del desenvolupament del projecte, sense cap mena de fi comercial o de distribució.

En cas de voler explotar les dades originals, cal remarcar que aquestes dades són propietat de les plataformes estudiades, i que en tot cas, sempre caldrà recurrir als correspondents termes fixats per els propietaris per tal d'obtenir el seu consentiment.

Codi

Els codis que hem fet servir per realitzar el Web Scraping el podran trobar al següent enllaç: https://github.com/jcarbonellsilva/web_scraping

Dataset

Els dataset el podran trobar als següents links:

- https://zenodo.org/record/6445962#.YlQBstPMKM8
- https://github.com/jcarbonellsilva/web_scraping

Taula de contribucions

Contribucions	Signatura
Investigació previa	Jordi Carbonell Silva , Carlos Solís García
Redacció de les respostes	Jordi Carbonell Silva, Carlos Solís García
Desenvolupament del codi	Jordi Carbonell Silva, Carlos Solís García