Pràctica 1 - Web Scraping

Tipologia i cicle de vida de les dades (UOC)

DATASET: Els 20 productes més venuts a Ziicube.

Índex de continguts

1 Context

Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Un dels problemes que té la venda online és que degut a la gran quantitat de productes que s'ofereixen en un mateix portal i la velocitat en que aquests canvien, costa poder tenir un control d'aquests productes i de com evolucionen.

Aquest treball s'ha centrat en la web del portal Ziicube (https://www.ziicube.com), que es dedica a vendre productes relacionats amb els cubs de Rubik, que a diari publica una llista dels 20 productes més venuts durant els últims 30 dies. És un tipus de web on la informació et desborda, la llista de productes que ofereix és molt gran i constantment sorgeixen noves incorporacions.

Per aquest motiu he trobat adient obtenir un dataset amb els 20 productes més venuts, perquè d'una forma ràpida pots saber com evolucionen les tendències dels productes.

2 Títol

Definir un títol que sigui descriptiu pel dataset.

Els 20 productes més venuts a Ziicube.

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 escollit.

El dataset conté una relació dels 20 productes més venuts, els últims trenta dies, a la web del portal de venda de cubs de Rubik ziicube.

4 Representació gràfica

Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.

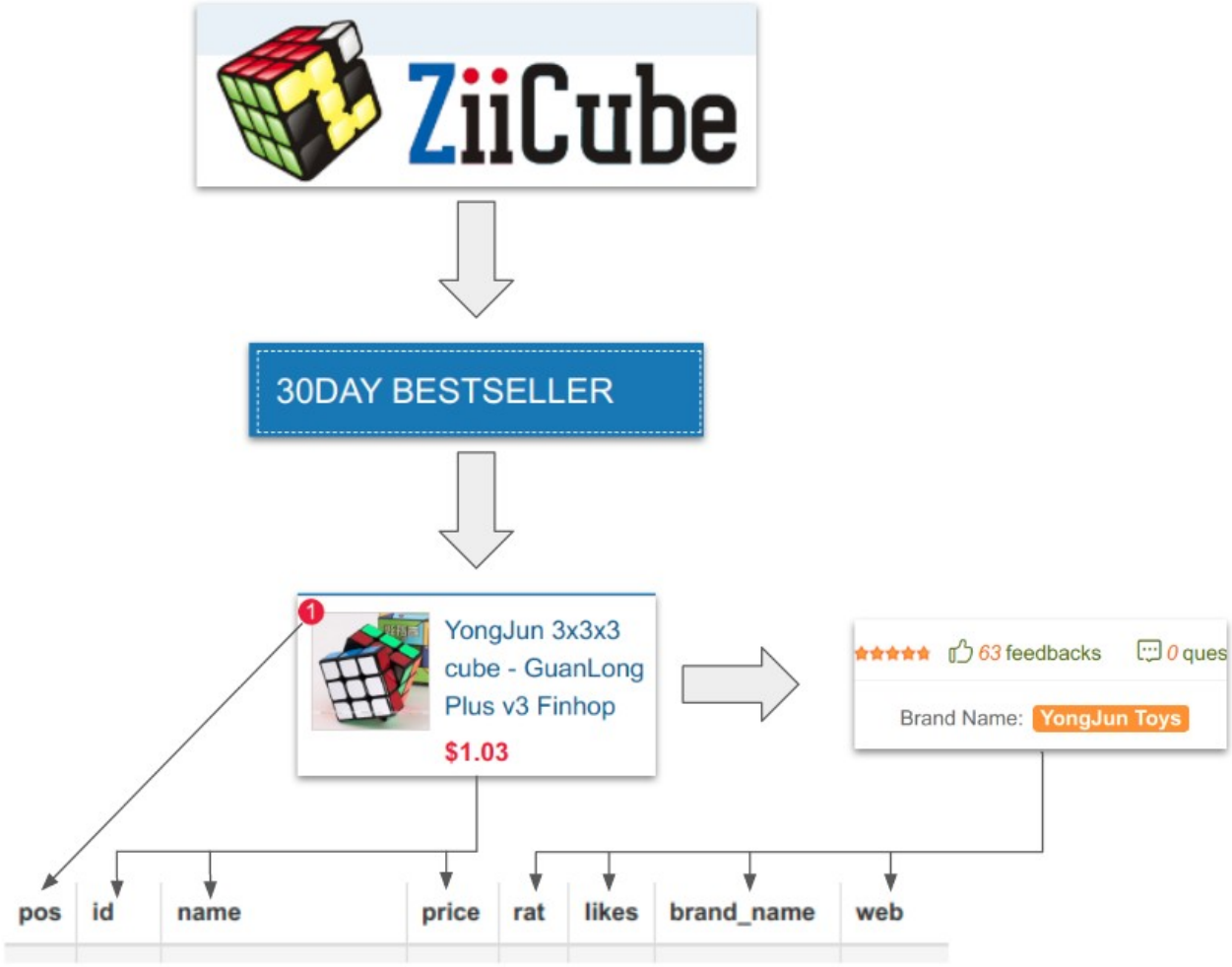


Figura 1: Esquema del dataset

5 Contingut

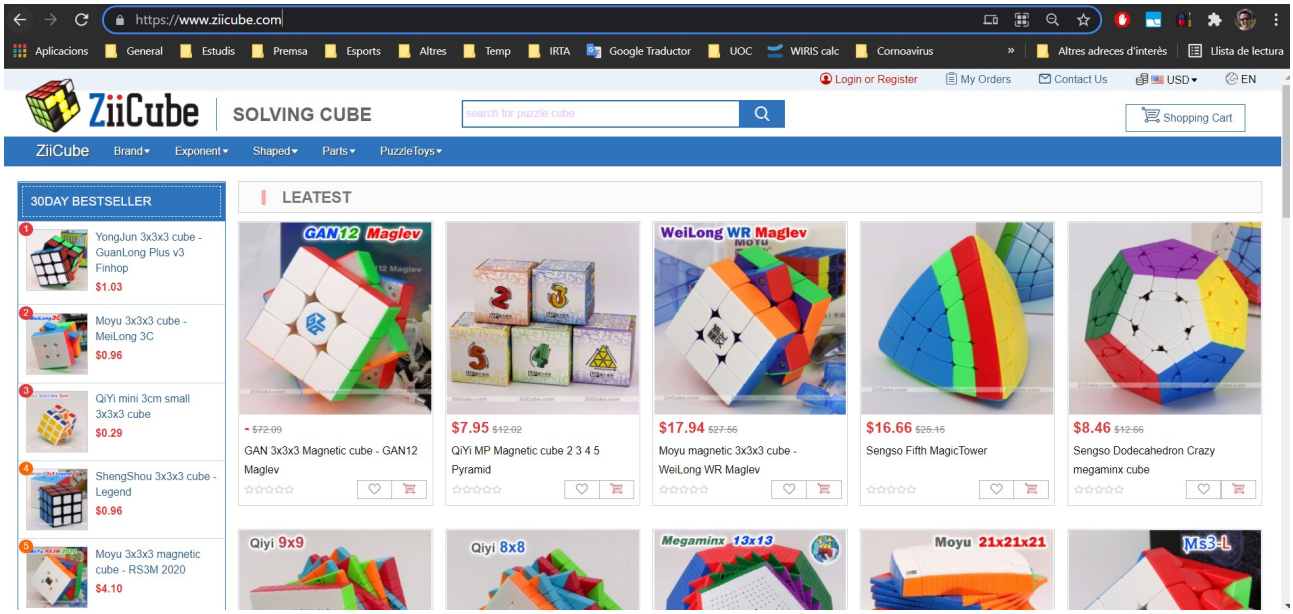
Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

Cada producte té els següents atributs:

- ***pos***: (int) Posició en el ràrquing dels més venuts
- ***id***: (str) Codi d'identificació del producte
- ***name***: (str) Nom del producte
- ***price***: (int) Preu del producte en $
- ***rat***: (int) Valoració del producte per part dels consumidors (0-100)
- ***likes***: (int) Número de m'agrades que ha rebut el producte
- ***brand_name***: Nom de l'empresa que produeix el producte
- ***web***: (str) Pàgina web del producte

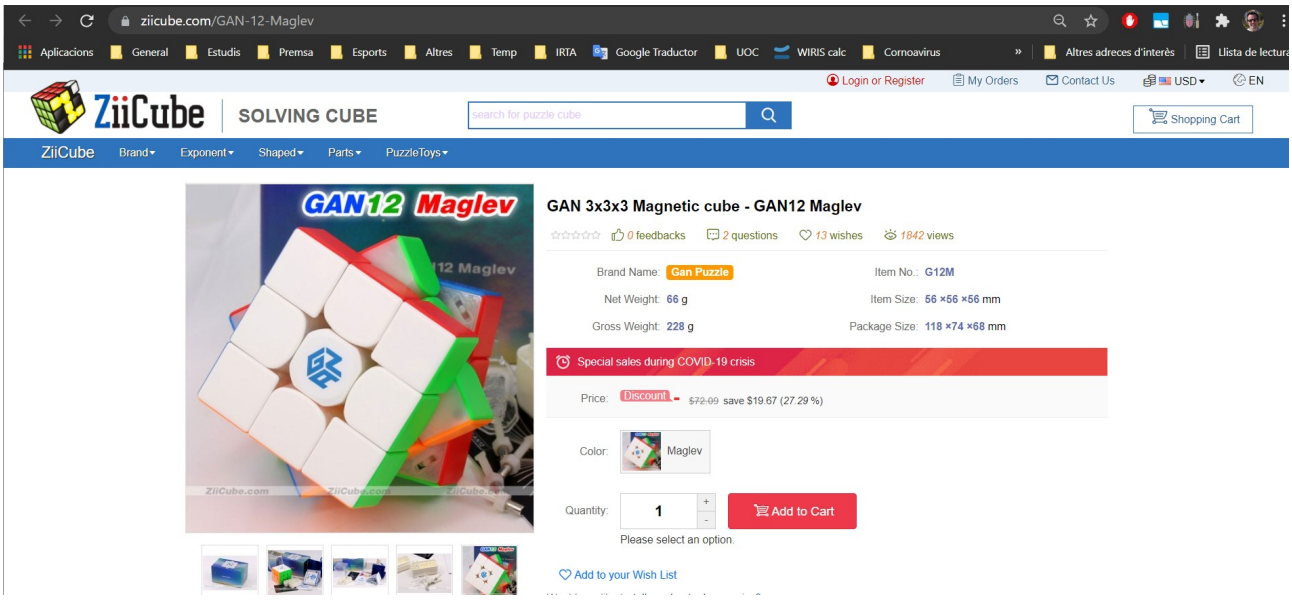
Les dades tenen un període de temps de 30 dies, és a dir que es computen les vendes dels últims trenta dies. S'actualitza cada dia.

La llista amb els 20 productes més venuts els últims 30 dies es troba a la part esquerra de la pàgina principal del portal, en un apartat anomenat '30DAY BESTSELLER'. Tal i com s'observa a la següent imatge (Il·lustració 1):



Il·lustració 1: Pàgina principal del portal ZiiCube

Cada producte conté un link que et porta a la pàgina web del producte en qüestió, tal i com s'observa a la següent imatge (Il·lustració 2):



Il·lustració 2: Pàgina d'un producte del portal ZiiCube

D'aquesta pàgina n'obtenim el nom de l'empresa (Brand Name), el percentatge d'estrelles (rat) i el número de feedbacks (likes).

6 Agraïments

Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

El propietari del portal ZiiCube és l'empresa *Zcube Limited to Deformation Century Trading Co., Ltd.*, ubicada a la Xina. Es dediquen a la venda de cubs de rubik i altres productes relacionats que hi tenen relació.

Per saber si a nivell legal podia extreu les dades del portal, he analitzant la seva llicència d'us[1]. Segons aquesta, no permet el seu ús per projectes comercials, però si pels que no tenen finalitats comercials. A més a més, la finalitat d'aquest projecte és facilitar el procés de decisió a l'hora de comprar un producte, fet que encara s'adequa més a la llicencia.

En relació a treballs anteriors, no n'he trobat cap que específicament es centrés a extreure els valors del portal ZiiCube, però si que n'he trobat molts relacionats en treure una llista de productes amb els seus preus d'un portal d'internet. Per exemple el repositori de GitHub anomenat Web Scraping Price[2], que t'ho permet descarregar de diferents portals com Amazona o Flipkart.

Un altre projecte similar és el repositori de GitHub Amazon_Best_Seller_Scraper[3], que et genera una llista setmanal dels llibres de no ficció més venuts a Amazon.

1 https://www.ziicube.com/index.php?route=information/terms_privacy
2 https://github.com/VipinindKumar/Web-scraping-price
3 https://github.com/Holly-Transport/Amazon_Best_Seller_Scraper

7 Inspiració

Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

La funció del conjunt de dades és captar quines són les tendències dins de les vendes de cubs de Rubik. La gran quantitat d'informació que dona el porta fa difícil veure d'una forma ràpida i concisa els productes més importants, fins i tot tenint-los en una part de pàgina principal. A mes a més aquesta llista evoluciona diàriament i això encara fa més difícil seguir-ne les evolucions. En canvi amb aquest projecte, cada dia pots tenir la llista actualitzada en qüestió de segons.

A la vegada, com que la llista hi ha una serie de variables, a part del nom, t'ajuden a poder entendre millor el producte. Ja sigui per la popularitat que pot tenir (ratio entre 'rat' i 'likes') o pel preu o per l'empresa que el fabrica. A més a més tens el link per poder accedir a la seva pàgina web.

Com que específicament d'aquest portal no hi havia cap anàlisis anterior, no podem fer una comparació directa. Però per exemple si ho comparem amb la el repositori de GitHub Amazon_Best_Seller_Scraper[4], hi hem afegit algunes variables que no estaven presents, com per exemple el preu o la web per poder-lo comprar. A la vegada la periodicitat també l'hem escurçat, de setmanal a diaria.

4 https://github.com/Holly-Transport/Amazon_Best_Seller_Scraper

8 Llicència

Seleccionar una d'aquestes llicències pel dataset resultant i justificar 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***

La llicencia seleccionada és la **CC BY_NC 4.0 Licence**, perquè crec que és la que millor s'adequa al projecte. Ja que permet que el dataset sigui utilitzat lliurement mentre sigui citat l'autor i mentre sigui en projectes sense ànim de lucre. Entenc que es interessant obrir el dataset a tothom que el vulgui fer servir, però en el cas que sigui per finalitats comercials les condicions haurien de ser diferents. Pel que fa al tema de citar a l'autor, també considero just que se'n recalqui l'autoria.

9 Codi

Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

https://github.com/magipamies/ziicube_scraping

10 Taula de contribucions

A més, al final del document, ha d'aparèixer la següent taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació per part del grup que l'integrant ha participat en aquest apartat. Tots els integrants han de participar en cada apartat, per la qual cosa, idealment, els apartats haurien d'estar signats per tots els integrants.

Contribucions	Signatura
Investigació prèvia	MPS
Redacció de les respostes	MPS
Desenvolupament del codi	MPS

11 Dataset

Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

https://doi.org/10.5281/zenodo.5644225