TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 1

Ferran Gurri Mancera

02/04/2018

Índex

1. Títol del dataset

Com a títol del dataset defineixo: “IT ebooks from Amazon”.

2. Subtítol

Com a subtítol: “Complete IT ebooks library from Amazon at march 2018”.

3. Imatge

Per representar el meu dataset he triat la següent imatge[1]:

He triat aquesta perquè volia una llibreria i l’efecte de llum sembla un passadís cap a la modernitat (ebooks).

4. Context

El dataset inclou informació de tots els llibres de la secció d’ebooks “Informatica, internet y medios digitales” del portal de venta online Amazon (https://www.amazon.es/). A partir de la informació obtinguda podem tenir una representació actual del catàleg de llibres que ofereix un dels portals de venta online més utilitzats.

5. Contingut

El dataset inclou els següents camps:

- title: títol del llibre
- prize: preu en euros
- ratings: nombre de valoracions que ha rebut el llibre
- stars: valoració promig en una escala del 0 al 5
- publish_date: data de publicació del llibre electrònic

[1] Imatge extreta de https://arena.westsussex.gov.uk/web/arena

- authors: llista d'autors

Les dades s'han obtingut en una sola descàrrega completa efectuada el dia 30 de Març del 2018 a les 00:30. Cal destacar que el portal presenta els seus productes paginats de 15 en 15 resultats i ha calgut fer un primer scraping per obtenir el total de pàgines a descarregar i adaptar l'script perquè dinamicament descarregui totes les pàgines disponibles. L'script està preparat per ser llençat més vegades sobrescribint el dataset i s'adaptarà al número de pàgines total de resultat.

6. Agraïments

Les dades pertanyen a Amazon.es i han sigut descarregades mitjançant el llenguatge de programació python i tècniques de web scraping per extraure informació a partir de documents HTML.

La idea inicial de la pràctica era crear un script de web scraping a algún buscador de vols i hotels però cap de les pàgines web que vaig provar ho permetien (skyscanner, booking, kayak, atrápalo, etc). Lo màxim que vaig arribar a aconseguir va ser una llista d'hotels del portal booking però sense preus. L'objectiu era programar l'script per llençar-lo durant 1 setmana a diferents hores i dies i analitzar la variabilitat dels preus dels vols i hotels a una ciutat específica.

Un cop descartada la idea inicial, vaig valorar la possibilitat de recollir dades d'algun repositori oficial que permetés web scraping sense restriccions[2], però he preferit buscar algo més proper a informació que em resultaria interessant per a mi i també, de passada, fer un anàlisi real de lo fácil/difícil que és trobar una página web que permeti ser rastrejada.

Amazon ho permet tècnicament (i no ho prohibeix explicitament en les seves normes d'us, en parlaré més en l'apartat de llicència) i em va semblar un portal amb molta informació disponible i d'interès general. L'altra portal que també em permetia web scraping era Wallapop però la vaig descartar per com gestionava les cerques per proximitat (no les inclou en la URL i dificultava l'scraping).

Com a conclusió personal després d'haver analitzat diferents portals, crec que aquells que basen el seu negoci en la publicitat present en les seves pàgines posen més recursos tècnics a l'hora d'evitar web scraping.

7. Inspiració

Amb aquest conjunt de dades podem:

[2] Vaig consultar el post https://towardsdatascience.com/scraping-the-internets-most-popular-websites-a4c6f0be382d i els llocs allà indicats com a "Complete Allow Sites"

- Buscar llibres de tecnología per nom d'autor, data de publicació, valoracions dels usuaris, nombre de valoracions o títol
- Analitzar els llibres més rellevants (amb més quantitat de valoracions) o els millor valorats
- Crear un mapa conceptual a partir de les paraules claus detectades al títol i analitzar d'aquesta manera quins són els temes de més actualitat
- Detectar autors amb més llibres publicats (caldria un tractament extra del camp "authors" ja que en el dataset és un String amb la llista de noms)

8. Llicència

A l'hora de triar quina llicència utilitzar he consultat la web https://help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-datasets

Crec que la llicència que millor s'ajusta a aquest dataset és la Creative Commons BY-NC-SA perquè:

- Permet fer modificacions i ampliacions del dataset per si algú vol mantenir la llista de llibres actualitzada o afegir-hi alguna columna
- Obliga a citar l'origen
- No permet us comercial de les dades (per no perjudicar al propietari de les dades: Amazon)
- Permet compartir qualsevol modificació del DataSet només amb la mateixa llicència que la utilitzada aquí. Evito així que directament o indirecta es perjudiqui a Amazon. Aquesta característica és amplicació del punt anterior

9. Codi

Codi font disponible en aquest mateix repositori de github.

10. DataSet

Codi font disponible en aquest mateix repositori de github.