Tipologia i cicle de vida de les dades

Arnau Rafi Cuello

Jordi Marsol López

Pràctica 1: Web scraping

Context

Aquest repositori es realitza en el marc de la Pràctica 1 de la assignatura *Tipologia i cicle de vida de les dades* del Màster en Ciència de Dades de la UOC. Consisteix en aplicar diferents tècniques i llibreries amb llenguatge de programació Python per tal d'extreure dades d'un lloc web mitjançant eines de *web scraping* i exportar-les a un dataset. En el nostre cas, extraurem dades del web https://rcdb.com (*Roller Coaster Data Base*), d'ara en endavant l'anomenarem *RCDB*.

Aquest lloc web és una base de dades que conté més de 10.000 muntanyes russes de tot el món amb les seves característiques. Ha estat elaborada de manera particular per diversos aficionats que la van actualitzant constantment i posa de manifest l'interès que pot tenir aquesta manera de divertir-se.

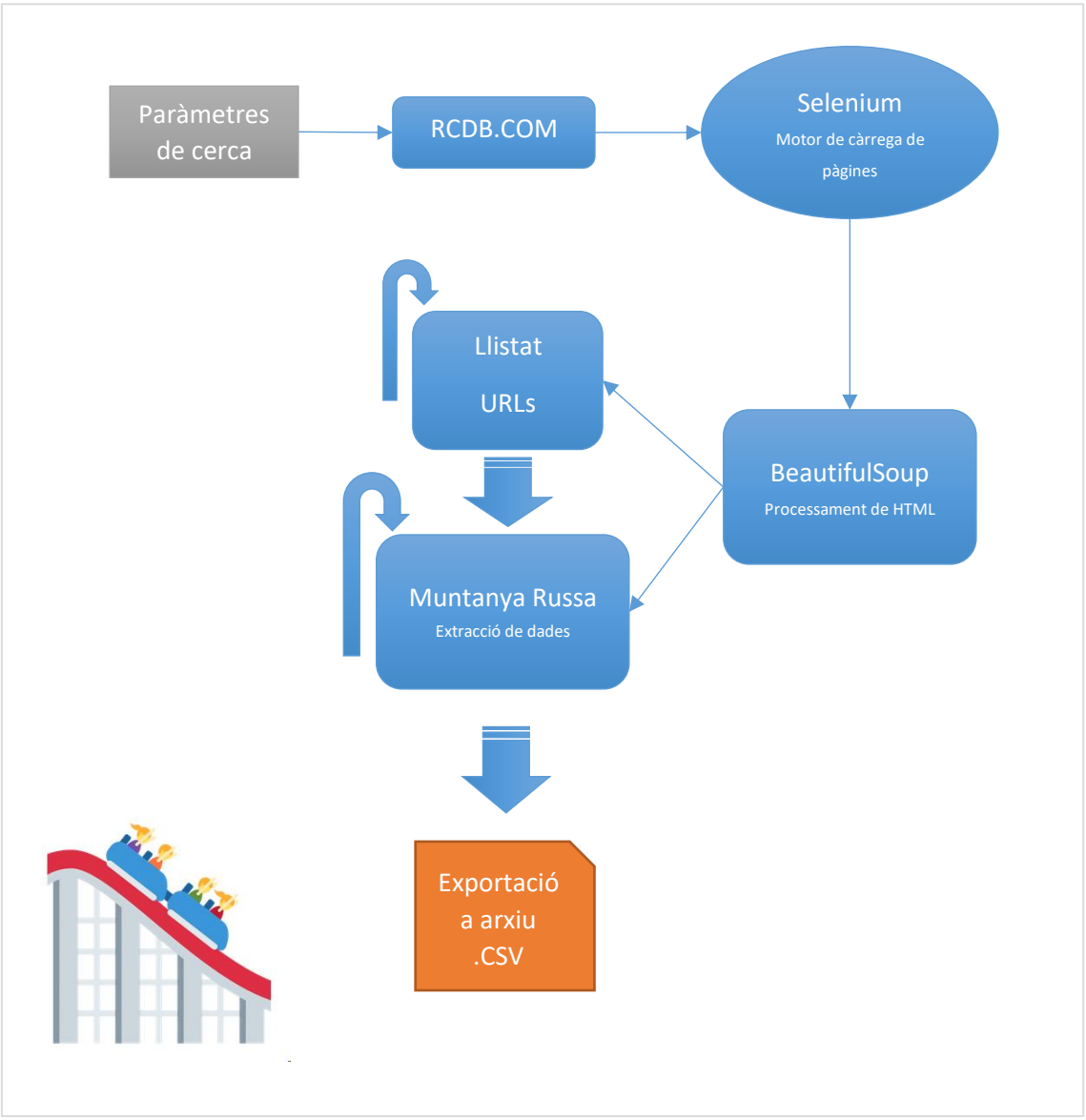
Títol del dataset

Muntanyes Russes operatives del web *Roller Coaster DataBase*

Descripció del dataset.

El dataset conté diverses dades referents a moltes de les muntanyes russes operatives que hi ha actualment al món. El lloc des d'on extraiem les dades contempla també muntanyes russes que ja no son operatives però aquí ens limitem a les que son vigents. Igualment donada la naturalesa educativa de la pràctica no pretén ser una extracció exhaustiva de les dades que conté el lloc i està limitada a un determinat nombre de característiques i a un nombre limitat de pàgines.

Representació gràfica



Contingut

Camps

- *Muntanya_russa*: Nom de la muntanya
- *Ubicacio*: Ubicació de la muntanya. Aquesta inclou diferents localitzacions com ara població, regió local i país, en aquest ordre normalment, però depèn del lloc on estigui en pot tenir més o menys.
- *parc*: Nom del parc d'atraccions on està ubicada.
- *Tipus*: Tipus de muntanya russa referit al material principal en què està fabricada.
- *Data_obertura*: La data d'inauguració de la muntanya.

- *Disseny*, *Fabricant*: El disseny i fabricant.
- *Model*: El model
- *Velocitat_màxima (mph)*: La velocitat màxima que pot assolir en milles per hora.
- *Llargada (ft)*, *Altura_màxima (ft)*: La llargada i l'altura màxima que té mesurada en peus.
- *Inversions*: Inversions (girs) que té la muntanya.
- *Duració*: Duració en minuts de tot el recorregut.
- *Elements*: Elements de què està composta la muntanya.

Periode de temps

El període de temps en que es genera el contingut del scraping és aquell en que s'executa l'aplicació. No hi ha la intenció de limitar la cerca en un període de temps determinat però si en el nombre de pàgines que s'extreuen. La web de rcdb.com actualitza freqüentment els canvis i noves muntanyes russes per tant la informació que s'extregui en executar l'aplicació suposem que serà el més actual possible.

Com s'ha recollit

El primer problema és que en fer un scraping amb la llibreria *requests* directament, la taula que mostra pel llistat, els *tags* no estan tancats i no es pot fer un bon seguiment dels elements, per això millor fem servir la llibreria *Selenium* que ens facilita la presentació de la jerarquia dels elements.

El lloc fa servir molt pocs identificadors únics per tant ens hem d'anar movent entre elements de manera relativa o en funció dels textos que hi apareixen.

L'altre problema és que no sempre surt tota la informació en tots els camps per tant en escanejar un valor pot ser que esperem que aparegui un però no hi sigui. Quan això passa omplim el valor amb el text *NA.*

Agraïments

RCDB és una gran base de dades de muntanyes russes d'arreu del món. Inclou informació detallada de milers d'aquestes màquines de diversió com per exemple: la seva altura, longitud, ubicació, velocitat màxima, fabricant, any d'inauguració, dissenyador...

Presenta una política de completa permissivitat amb bots, segons indica el seu robots.txt, cosa que agraïm a l'hora de fer *web scraping* :

```
User-agent: *
Disallow:
```

Hem pogut trobar un projecte de *scrape* similar de muntanyes russes nord-americanes (https://github.com/aarmora/jordan-scrapes-rcdb) però està exportat a .json.

Donat que es tracta d'un lloc no comercial no voldríem abusar dels recursos del lloc web per fer la extracció. Creiem que si se'n volgués fer una extracció completa possiblement per algun us més comercial seria preferible contactar amb els autors de la pàgina. Com indica als termes d'us:

Using the content to construct other databases, websites or applications requires prior written permission.

Inspiració

Tot i ser una base de dades bastant extensa, no disposa d'un API per a descarregar les dades ni d'eines per fer anàlisis profunds amb les seves dades. És per això i sumat el fet que està plenament codificat en HTML, que ens ha semblat un bon web des d'on iniciar i practicar el *web scraping*.

Intentarem crear un dataset de muntanyes russes globals des del que es pugui fer anàlisis profunds de les muntanyes russes que hi ha al món.

A partir d'aquest dataset es podrien respondre, per exemple, les següents preguntes:

- *Quina és la evolució de la altura (o longitud, o velocitat màxima...) de les muntanyes russes al llarg de la història?*
- *Quins dissenyadors construeixen muntanyes russes més altes (o llargues, o ràpides...)?*
- *Si es vol contractar una constructora de muntanyes russes, quina és la empresa que té més experiència en muntanyes russes de fusta?*

Entre d'altres...

Llicència

Es farà servir una llicència pels datasets de tipus "Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International" (CC BY-NC-SA)

La motivació del seu us ve pel fet principal que l'scraping s'ha fet a partir d'un interès educatiu i per tant no es permet un us comercial d'aquest. També es requereix la citació dels autors originals i que qualsevol transformació o addició del codi original ha de seguir tenint el mateix tipus de llicència

Codi

El codi es pot trobar al lloc https://github.com/jordi-marsol/rcbd_webscraping

Dataset

Una còpia del dataset es pot obtenir a https://zenodo.org/record/5639556

Participants

Contribucions	Signatura
Investigació prèvia	ARC,JML
Redacció de les respostes	ARC,JML
Desenvolupament del codi	ARC,JML