Pràctica 1: Web scraping

Descripció

Aquesta pràctica s'ha realitzat per a l'assignatura de Tipologia i cicle de vida de les dades, del Màster en Ciència de Dades de la Universitat Oberta de Catalunya. Es fa ús de tècniques de web scraping per a extreure dades de pel·lícules en la cartelera de Filmaffinity i generar un dataset.

Membres de l'equip

La pràctica ha sigut realitzada per Josep Tormo Costa i Oriol Bardés Robles.

Fitxers del codi font

main.py: programa de scraping. movies_df.csv: fitxer csv amb el dataset resultant. movies_scrape/movies_df.csv: fitxer csv checkpoint.

Qüestions de l'enunciat de la pràctica

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

L'objectiu de la nostra pràctica és poder recollir programàticament un llistat de les pel·lícules disponibles en la cartellera de les sales de l'estat espanyol. Com a font de dades triem la secció de cartellera del portal FilmAffinity (www.filmaffinity.com)

Aquest portal no diposa de cap API de la qual en tinguem coneixement, de forma que és un bon cas d'ús per aplicar web scraping.

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

"Cartellera de pel·lícules als cinemes de l'estat espanyol"

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).**

Aquest dataset està composat pel llistat les pel·lícules actualment en cartellera, incloent alguns dels paràmetres més descriptius de cada pel·lícula com ara el títol, l'any,la durada i el país. També considerem interessant tenir la capacitat d'endreçar les pel·lícules de la cartellera en base a la seva valoració, per la qual cosa el dataset també recull el rating proporcionat per FilmAffinity.

Aquest llistat a dia 31-10-2020 està composat per 64 títols.

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

Com a esquema del dataset es llista a continuació les 5 primeres entrades, juntament amb la capcelera:

,title,year,duration,country,rating

0,Cars 3,2017,109 min., Estados Unidos,"5,9"

1,Trolls World Tour aka,,2020,90 min.,"5,6"

2,The Secret: Dare to Dream,2020,107 min., Estados Unidos,"4,8"

3,Divorce Club,2020,108 min., Francia,None

4,Hope Gap,2019,100 min., Reino Unido,"6,2"

5,Promare: Puromea,2019,111 min., Japón,"6,6"

6,Tuntematon mestari aka,,2018,94 min.,”7,1"

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

El dataset no està dissenyat per acumular registres durant un període de temps determinat sinó que es genera dinàmicament en cada execució del codi. Inclou els següents camps:

title: títol de la pel·lícula (e.g.”Cars 3”) year: any d’estrena en format yyyy (e.g. “2019”) duration: durada en minuts del metratge (e.g.”107 min”) country: país d’origen en format text (e.g. “Estados Unidos”) rating: valoració de 0 a 10 en format decimal FilmAffinity de la pel·lícula (e.g “5,6”)

Les dades s'han recollit usant les capacitats del paquet Selenium de Python per extreure les dades del portal en qüestió.

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

El conjunt de dades s’ha obtingut del portal www.filmaffinity.com, que és un lloc web espanyol dedicat al món del cinema. Va ser

creat an Madrid al maig de 2002 pel crític de cinema Pablo Kurt Verdú Schumann i pel programador Daniel Nicolás.

Cal destacar que no hi ha cap menció en les condicions d'ús del portal (https://www.filmaffinity.com/es/private.php) a pràctiques de web scraping.

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

Aquest conjunt de dades és interessant ja que permet capturar la informació de la cartellera en un moment donat en el temps, i permetre'n la seva anàlisi posterior. Un cop generat el dataset, també és possible creuar la informació extreta de FilmAffinity amb la informació proporcionada per altres portals similars (e.g www.imdb.com) usant tècniques s'scraping o APIs, permetent completar la informació o fitxa de cada pel·lícula.

Cal remarcar de nou que aquest portal no disposa de cap API per la qual cosa usar tècniques de web scraping ens sembla particularment adequat.

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

La llicència escollida pel dataset és CC0 (Public Domain) pels següents motius. Els autors del dataset no requerim ser mencionats ni volem haver d'autoritzar especificament el seu ús. Per tant, com que aquest dataset no ha de comportar mai cap obligació legal per a qui l'utilitzi i ha de ser utilitzat lliurement per tothom que ho vulgui, CC0 ens sembla l'opció més adequada.

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

El codi Python per a generar aquest dataset està disponible a:

https://github.com/jotorcos/web_scraping

10. **Dataset. Publicar el dataset en format CSV a Zenodo (obtencio del DOI)**

El dataset resultant s'ha publicat a Zenodo amb el següent DOI:

10.5281/zenodo.4165158

En el dataset els autors apareixen com jotorcos i obr-uoc, usant els usuaris guthub dels integrants de les pràctiques.

Contribucions

Contribucions	Signa
Recerca prèvia	Josep Tormo Costa i Oriol Bardés Robles
Redacció de les respostes	Josep Tormo Costa i Oriol Bardés Robles
Desenvolupament codi	Josep Tormo Costa i Oriol Bardés Robles