Universitat Oberta de Catalunya

uoc.edu

Pràctica 1: Anàlisi de contingut en Filmin mitjançant Web Scraping

Víctor Caballero

Pol Roselló Boqué

Tipologia i cicle de vida de les dades

Universitat Oberta de Catalunya

1. Context

Per dur a terme aquest projecte, s'ha recollit informació a través del web scraping de la pàgiona web de Filmin (https://www.filmin.es/catalogo) per obtenir un dataset que contingui informació sobre els idiomes tant de l'àudio com dels subtítols de totes les pel·lícules i sèries disponibles en el catàleg. El lloc web triat és adequat per a aquesta finalitat, ja que ens proporciona informació sobre totes les versions disponibles per a cada una de les pel·lícules i sèries, i també perquè les dades estan estructurades de manera que es poden extreure fàcilment mitjançant tècniques de web scraping.

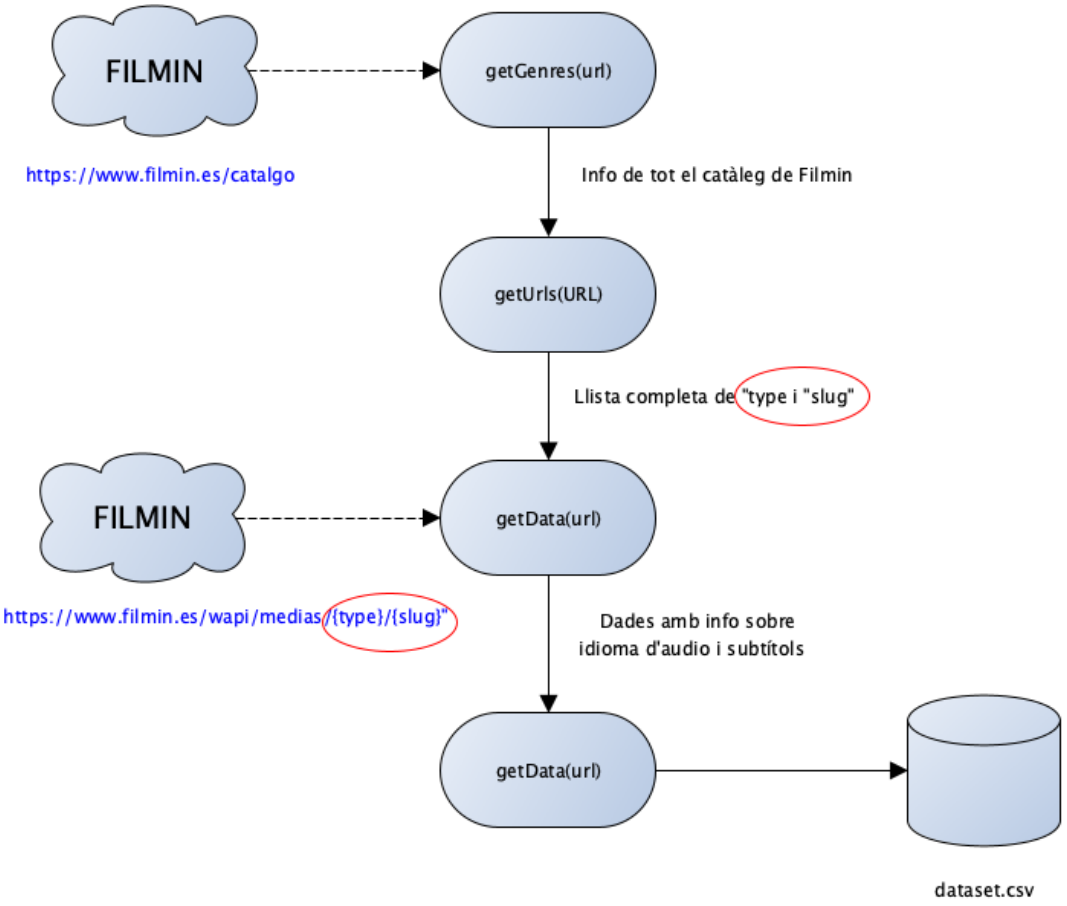
2. Títol

"Idiomes disponibles al catàleg de Filmin".

3. Descripció del dataset

El dataset extret consisteix en un arxiu csv que inclou informació sobre els idiomes de les pel·lícules i sèries disponibles al catàleg de Filmin. Així, a partir del dataset es podrà analitzar quantes pel·lícules i sèries estan disponibles en català al catàleg de Filmin.

4. Representació gràfica



5. Contingut

Title: títol de la pel·lícula o sèrie.

Type: tipus de contingut, en aquest cas totes les entrades són de tipus "pel·lícula" o "sèrie".

Year: any de producció de la pel·lícula.

Actors_info: informació sobre els actors principals de les pel·lícules

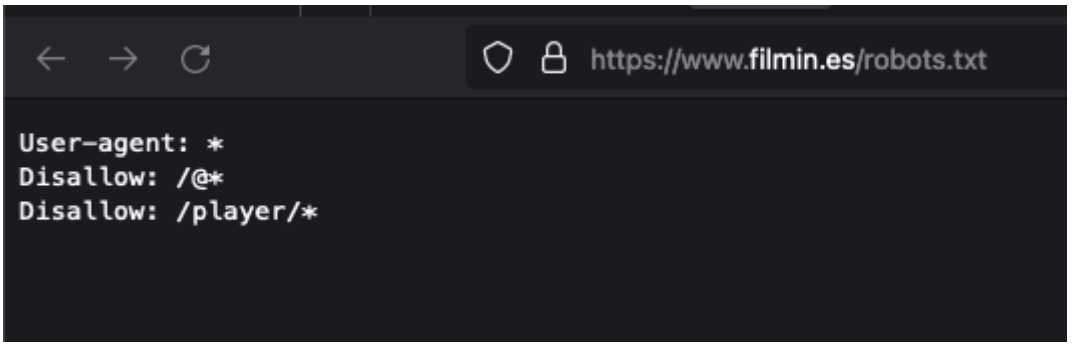
Language: informació sobre els diferents idiomes en què estan disponibles les pel·lícules i sèries del catàleg.

Subtitles: idiomes dels subtítols disponibles per a la pel·lícula.

6. Propietari

El propietari de les dades és Filmin, ja que hem recopilat les dades del seu lloc web mitjançant tècniques de web scraping. No es coneixen anàlisis anteriors per a aquest dataset en particular, però és important destacar que el web scraping ha estat realitzat de manera ètica i legal, respectant les polítiques del lloc web de Filmin i sense infringir cap llei de protecció de dades. Hem verificat els termes i condicions del lloc web per assegurar-nos que complim les seves normes. A més, hem respectat els drets d'autor de les obres que hem utilitzat en el projecte i hem assegurat que estiguin autoritzades per ser compartides en aquesta plataforma.

Així mateix, hem tingut en compte les restriccions imposades pel fitxer robots.txt de Filmin, que indica quines parts del lloc web són accessibles per als robots de cerca i quines no ho són. Hem evitat accedir a les seccions restringides i hem respectat les directrius del fitxer.



User-agent: * Aquest símbol ens indica que les regles que hi ha sota s'apliquen a tots els usuaris, es a dir, a tots els robots de cerca que accedeixin al lloc web
Disallow: /@* Aquest símbol ens indica que els robots de cerca no han d'accedir en qualsevol URL que comenci amb un /@ seguit de qualsevol caràcter. Un exemple podria ser un @usuari.
Disallow: /player/*: Aquesta regla prohibeix als robots de cerca accedir o indexar qualsevol URL que comenci amb "/player/" seguit de qualsevol caràcter. Per exemple, una URL com a "/player/123" estaria bloquejada per als robots.

En general, creiem que hem actuat de manera ètica i legal en tot moment en el context del nostre projecte a Filmin.

7. Inspiració

Aquest conjunt de dades proporciona una gran quantitat d'informació valuosa per a diversos públics, des dels amants del cinema fins als investigadors. Per als amants del cinema en català, aquesta informació pot ser una eina útil per descobrir noves pel·lícules i sèries en aquest idioma i explorar el catàleg de Filmin. També pot ser útil per als investigadors del cinema, la sociologia o la lingüística per entendre com es distribueixen les pel·lícules en diferents idiomes i com ha evolucionat aquesta distribució amb el temps.

A través dels camps del dataset, podem plantejar preguntes interessants com ara quines són les pel·lícules més populars en català, quins són els gèneres de sèries més comuns en català, i com ha evolucionat la distribució de pel·lícules en català al llarg del temps. Així mateix, podem utilitzar aquesta informació per investigar com ha canviat la preferència del públic per les pel·lícules en català al llarg de les dècades, i si hi ha una tendència creixent o decreixent en la demanda de contingut en aquest idioma. També podem explorar les diferents combinacions de llenguatges i subtítols disponibles per comprendre millor les preferències dels espectadors. En resum, aquest conjunt de dades té el potencial d'oferir una gran quantitat de coneixements valuosos per als diferents públics.

8. Llicència

Per a aquest dataset es selecciona la llicència CC BY-NC-SA 4.0, aquesta llicència és una opció adecuada per al nostre dataset, ja que permet compartir, copiar i redistribuir les dades, així com adaptar-les i reutilitzar-les, sempre que sigui amb fins no comercials i citant la font original.

9. Codi

El nostre codi conté diverses funcions que permeten obtenir informació sobre pel·lícules i sèries del lloc web Filmin.

1. La primera funció getGenres(url) s'encarrega de descarregar el catàleg complet de Filmin, però sense informació sobre l'idioma disponible. Per fer això, utilitza el mòdul urllib.request i afegeix múltiples capçaleres HTTP per simular una sol·licitud de navegador real. Es fa ús de la biblioteca brotli per descomprimir la resposta i es guarda en format JSON.
2. La funció getUrls(URL) s'encarrega d'obtenir les parts de la URL que canvien en cada contingut (slug i type). Utilitza un bucle while per simular el desplaçament a la pàgina web i anar canviant el paràmetre page en cada iteració. A continuació, guarda la informació en un DataFrame de pandas.

3. La funció getData(url) és similar a getGenres(url), però rep altres cookies en la sol·licitud. Aquesta funció s'encarrega d'obtenir informació addicional de cada pel·lícula o sèrie.
4. La funció getCsv(URL) és la funció principal que combina tot el procés. Extreu la informació rellevant del JSON obtingut, com ara títol, any, tipus de contingut, actors, àudio disponible i subtítols disponibles. Després, guarda la informació recopilada en un fitxer CSV.

Entre les dificultats que presenta el lloc web de Filmin es troben:

- La necessitat de simular una sol·licitud de navegador real mitjançant l'ús de múltiples capçaleres HTTP i cookies.
- L'estructura del catàleg, que requereix iterar sobre les pàgines i extreure les URL de cada contingut.
- La informació sobre idiomes i subtítols es troba en llistes niades en el JSON, la qual cosa requereix iteracions addicionals per extreure-la.

Per resoldre aquestes dificultats, el codi proporcionat implementa tècniques de web scraping, com l'ús de capçaleres HTTP personalitzades i cookies per simular sol·licituds de navegador, i la iteració a través de pàgines i llistes niades per extreure informació rellevant.

10. Dataset

Enllaç DOI: https://doi.org/10.5281/zenodo.7847618

11. Vídeo

https://drive.google.com/file/d/1XgP3uy2Vs7dtEoihocdSgCL8dTXAK09o/view?usp=share_link