Pràctica 1

Membres del grup:

- Aarón Puche Benedito
- Roger Pardell Carrera

1. Context

En un moment de pandèmia global, on molts governs han trobat en el confinament i les quarantenes la solució temporal a la situació, l'aturada de les interaccions socials ha provocat que l'oci s'hagi de consumir a casa i en petita escala. Molta gent ha trobat el passatemps perfecte en plataformes de pel·lícules en streaming, les quals proporcionen una quantitat de llargmetratges considerables. El problema sorgeix, masses vegades, en triar quin títol volem veure.

De la necessitat de tenir un document amb els noms de pel·lícules sorgí l'interès per a fer un Web Scraping d'una pàgina web amb aquesta informació: The Movie Database. Aquesta pàgina ens proporciona el conjunt d'informació que ens interessa: títol, any d'estrena, gènere, sinopsis, etc. Així doncs, podem proporcionar un llistat únic de tots els films fins ara rodats.

Tot i que pot semblar un argument banal, aquestes dades poden tenir un important valor per analitzar els gustos dels consumidors i productors al llarg dels anys els quals, barrejats amb elements històrics, es podrien treure conclusions de caràcter sociològics importants.

D'altra banda, aquest conjunt de dades podria ser fàcilment incorporat en aplicacions de caràcter comercial. Podria ajudar-nos a escollir la plataforma de vídeos en streaming que reprodueix més títols del nostre interès, o bé proporcionar les preferències dels consumidors respecte els gèneres o les sinopsis.

Volem fer menció que l'script carregat és per usuari d'Apple. En cas de voler usar-lo per Windows, s'hauria d'actualitzar el path del webdriver de Chrome.

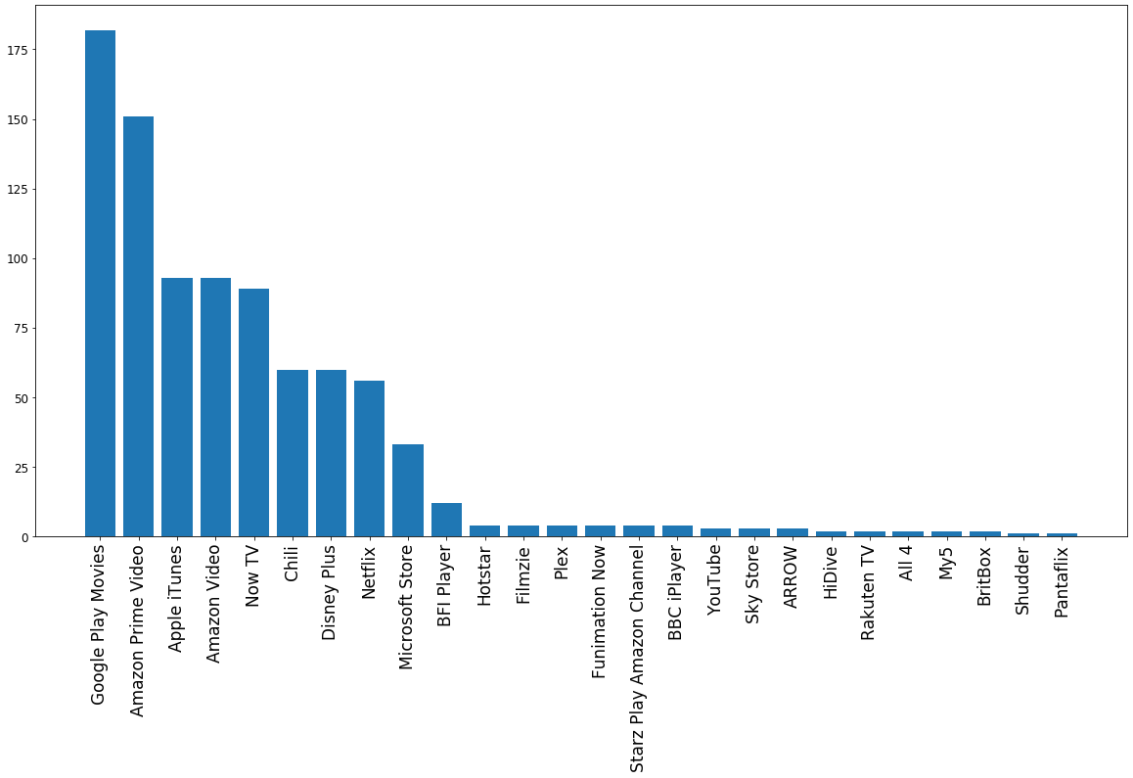
2. Títol del dataset

Pel·lícules d'acció estrenades entre el 2000 i el 2005.

3. Descripció del dataset

El data set resultant és compost de 1880 observacions, corresponent a totes les pel·lícules d'acció estrenades entre el 2000 i el 2005, i 12 atributs. Aquests atributs són:

4. Representació gràfica



En aquest gràfic tenim representada de forma visual la distribució de les pel·lícules segons la plataforma de streaming a la qual pertanyen, si és que hi pertanyen. Això vol dir que tot i que Google Play Movies és la plataforma amb més pel·lícules amb els filtres establerts, hi ha films que actualment no es troben en streaming. De fet, si sumem els valors de l'histograma, el resultat és de 874, lluny dels 1880. Per tant, 1006 títols de pel·lícula no es troben en cap plataforma.

5. Contingut

Com el títol especifica, aquest data set mostra diversos atributs de totes les pel·lícules d'acció estrenades entre el gener del 2000 i el gener del 2005. Cada observació consta de 12 atributs, especificats seguidament amb un exemple de "El Senyor dels Anells":

- Títol nom de la pel·lícula en castellà d'Espanya, el qual seria "El señor de los anillos".
- Títol original, el qual seria "The Lord of Rings".
- Idioma original, que seria l'anglès.
- Duració, en el format "hores minuts". En aquest cas, s'hauria de veure quin film de la saga. Tot i això, l'exemple seria "2h 59m".
- Gèneres, a més del d'acció. L'exemple triat té, a més a més, "aventures" i "fantasia".
- Directors, en el cas seleccionat, el director és Peter Jackson.
- Sinopsis.
- Data d'estrena, en format dd/mm/yyyy. El film d'exemple fou estrenat el 19 de desembre del 2001.

- Classificació per edat. Pot anar de U (permès en totes les edats) a 18 (permès a majors de 18 anys). En el cas del Senyor dels Anells, aquesta té una classificació PG (de control parental).
- Enllaç a la imatge de cartell. En aquest cas, hem mantingut aquesta variable per si volguéssim crear una app.
- Puntuació obtinguda pels visitants de la pàgina web. El nostre exemple té una puntuació de 83 sobre 100.
- Plataforma de streaming on podem trobar-la. Finalment, podem trobar la pel·lícula a Now TV.

La recollida de la informació s'ha efectuat amb Python, utilitzant l'eina Selenium. Se li ha donat l'enllaç de la pàgina on s'exposa el conjunt de pel·lícules i, a partir d'allà, s'han seleccionat els filtres. Acte seguit, el codi estén la pàgina fins a carregar el màxim de pel·lícules amb aquelles característiques. Després, es guarda en una llista les terminacions de totes les extensions de les pel·lícules. Amb el llistat, el codi carrega la pàgina de cada film i n'extreu les dades rellevants per el nostre projecte. Aquestes són emmagatzemades en un fitxer csv.

6. Agraïments

Hem decidit extraure informació de la web The Movie Database.

Aquest lloc pertany a TiVo Corporation i és un projecte que va començar en 2008 on una comunitat de gent començà a agregar i editar informació sobre pel·lícules i series. Com diuen ells, volien crear una Wikipedia molt especialitzada.

A partir de les dades d'aquesta web s'han creat algunes aplicacions. Podem veure algunes de les més destacades en aquest enllaç. Ens ha cridat l'atenció l'app de Cinematics, disponible a Android. Ens permet veure informació sobre pel·lícules i sèries d'una manera senzilla i molt fluida.

Hem revisat si disposava de l'arxiu robots.txt per no el te.

https://www.themoviedb.org/robots.txt

7. Inspiració

En aquesta web podem aconseguir un conjunt de dades molt interessant. En permet tant fer anàlisis com crear aplicacions a l'estil de l'app de Cinematics.

Pel que fa a les anàlisis, podríem veure com evolucionen els distints gèneres llarg del temps. Hi ha un gènere que destaca en una franja d'anys concreta? També podem veure la situació de les plataformes streaming. Hi ha alguna plataforma que sigui més predominant? A partir de quin any comencen a tindre un volum significant de pel·lícules?

Per l'altra banda, podem realitzar alguna aplicació de seguiment de pel·lícules. A més, a partir de les dades dels usuaris que utilitzaren la nostra web podríem crear un sistema de recomanació i mostrar de forma més precisa les pel·lícules més interessant que els hi puguin agradar.

8. Llicència

Hem decidit que aquest dataset estigui baix la llicència de Public Domain License. Aquesta llicència renuncia a qualsevol dels drets al fitxer se'ns pogués atorgar pel fet de ser-ne els autors. Per tant, tothom és lliure de descarregar-lo, copiar-lo, modificar-lo, distribuir-lo i interpretar-lo sense demanar permís. De fet, també es permeten les accions anteriors amb propòsits comercials.

Per exemple, aquesta llicència permetria a un estudiant d'universitat fer anàlisis propis sobre com ha evolucionat la producció de pel·lícules al llarg dels anys o analitzar l'impacte de les plataformes de streaming.

També permetria a emprenedors/es crear aplicacions pròpies amb caràcter lucratiu a partir del data set sense cap limitació ni legal ni monetària.

9. Codi

Codi en el Jupyter Notebook "Obtenir dataset pelicules.ipynb"

10. Dataset

Dataset disponible a Zenodo a través d'aquest enllaç.

DOI 10.5281/zenodo.4647054

Contribucions	Signa
Recerca prèvia	Aarón, Roger
Redacció de les respostes	Aarón, Roger
Desenvolupament codi	Aarón, Roger