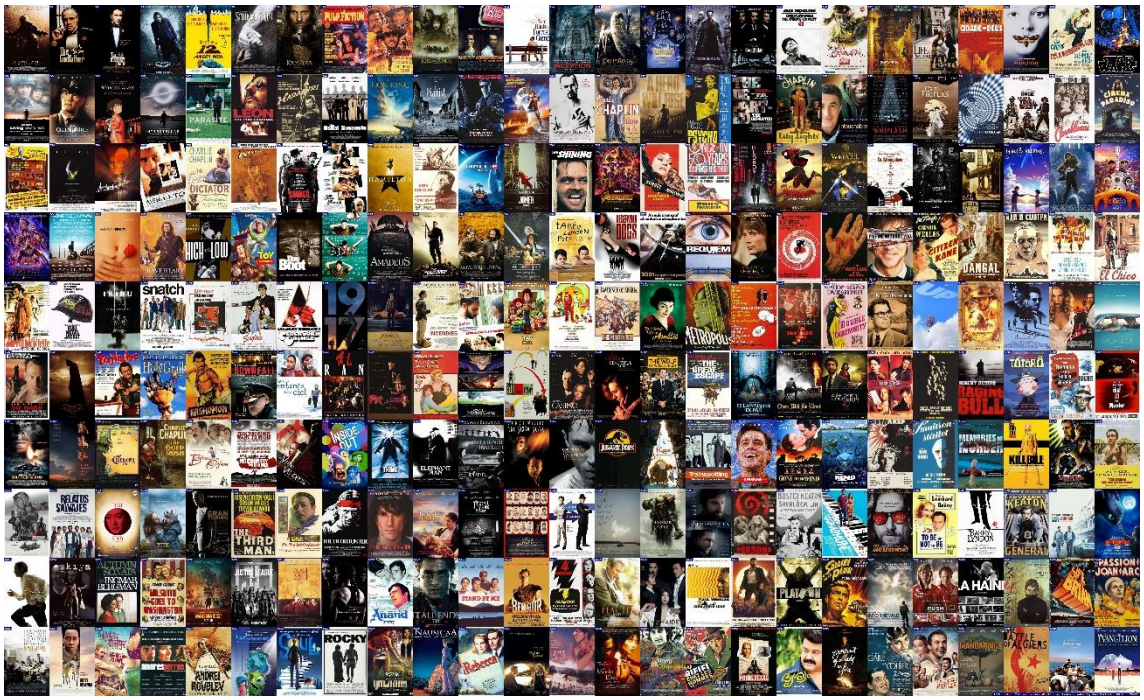
Tipologia i Cicle de vida de les dades

Pràctica 1 – Web Scraping

Autor: Héctor Gutiérrez Muñoz

Dataset: **Dades tècniques de les pel·lícules al top 250 d'IMDb**



Imatge: Macro-pòster amb els pòsters de les 250 pel·lícules més rellevants a IMDb.
Font: https://movies.luka.in.rs/poster

Descripció

Aquest dataset recull informació tècnica sobre les pel·lícules que actualment (13 d'abril de 2021) són considerades com les millors de la història per part de la comunitat d'usuaris sobre cinema més gran i important d'Internet, la IMDb (Internet Movie Database). Al dataset no hi ha informació sobre actors o directors, sinó que ens centrem en aspectes més tècnics, com la duració, el pressupost o la data d'estrena, dades a partir de les quals podrem fer una anàlisi posterior.

Context

Les pel·lícules que hi apareixen al dataset són de tota mena, tant velles com modernes i en diversos llenguatges. Les dades s'han recollides d'IMDb no només pel fet que és la base de dades

més completa que podem trobar a Internet, sinó també perquè la seva comunitat és molt gran i activa. Així doncs, la mostra de 250 pel·lícules que hi recollim és la que millor podem acceptar com a llista de les pel·lícules més importants de la història i tenim les pel·lícules puntuades, la qual cosa és important per respondre a les preguntes que formularem a continuació.

Contingut

Com ja s'ha dit, les dades que s'hi recullen de les pel·lícules són més tècniques i apropiades per a una anàlisi posterior. Per a cada pel·lícula disposem de les següents dades:

- **Name:** Nom de la pel·lícula.
- **Score**: Puntuació mitjana dels usuaris d'IMDb.
- **Genre1**: Gènere més important de la pel·lícula (IMDb assigna a cada pel·lícula fins a 3 gèneres diferents).
- **Genre2**: Segon gènere més important de la pel·lícula (IMDb assigna a cada pel·lícula fins a 3 gèneres diferents).
- **Genre3**: Tercer gènere més important de la pel·lícula (IMDb assigna a cada pel·lícula fins a 3 gèneres diferents).
- **Duration**: Duració de la pel·lícula en minuts.
- **Release**: Data d'estrena de la pel·lícula (formatejat com *13 April 2021*).
- **Rating**: Qualificació d'edat de la pel·lícula als EUA.
- **Country**: País de rodatge de la pel·lícula.
- **Language**: Llenguatge principal de la pel·lícula.
- **Sound**: Tipus de tecnologia de so fet servir al rodatge (normalment n'hi ha més d'un, separats per |).
- **Color**: Color o B/N.
- **Ratio**: Relació d'aspecte de la pel·lícula (en format X : 1).
- **Budget**: Pressupost de la pel·lícula en la moneda del país sense ajustar a la inflació (format $100,000,000 o GBP100,000,000, amb codi ISO en cas que la moneda no sigui dòlar).
- **Gross**: Recaptació mundial de la pel·lícula en dòlars sense ajustar a la inflació (format $100,000,000).
- **BadReviews**: Nombre de valoracions d'usuaris amb una puntuació entre 1 i 4.
- **NeutralReviews**: Nombre de valoracions d'usuaris amb una puntuació entre 5 i 7.
- **GoodReviews**: Nombre de valoracions d'usuaris amb una puntuació entre 8 i 10.

Com ja s'ha esmentat, a dins d'aquest top 250 hi trobem pel·lícules de tota mena. Aquest top, que s'ha recollit a 13 d'abril de 2021, no és fix, sinó que de fet evoluciona amb el temps, ja que la comunitat IMDb cada dia és més activa. Tot i això, gairebé mai no s'hi produeixen grans canvis al top, per la qual cosa podem afirmar que la mostra que recull el dataset és representativa de les pel·lícules més importants de la història segons el públic general.

Agraïments

He d'agrair a l'equip que és al darrere de la web IMDb per haver creat una base de dades sobre cinema molt gran, acurada i (almenys en part) lliure. Però no només a ells, sinó que també s'ha d'agrair a tots els usuaris, professionals i aficionats, que fan servir aquesta web i l'han convertit en la comunitat cinèfila més important d'Internet.

També agraeixo a les altres persones que abans d'aquest treball han fet les seves anàlisis sobre aquestes pel·lícules amb dades d'IMDb, Són molt interessant les anàlisis de Max Tohline (de Bright Lights Film)[1] o Mark Lee (d'Overthinking It)[2] sobre com ha canviat aquest Top 250 i com això reflecteix el canvi en la nostra societat. Finalment, agrair una anàlisi més senzilla i estàtica com la de Mukul Jain a Medium[3], però que fa servir Web Scraping i és la que més va inspirar-me a fer aquesta feina.

Inspiració

La inspiració per extreure aquest conjunt de dades ha estat completament personal, ja que el cinema és la meva passió més gran. Com que realment, més enllà de l'art, el cinema representa una indústria molt gran, crec que analitzar les pel·lícules millor considerades pel públic general pot ser molt interesant, tant a dins de la indústria com pels aficionats.

Més concretament, amb aquest dataset les preguntes que s'intenten respondre són preguntes relacionades amb el gust cinematogràfic de la gent (per això hem inclòs les dades sobre les valoracions dels usuaris), però no tant relacionant aquest gust amb els actors, els directors i en general, la part més artística i social (com fan els tres articles citats abans), sinó que aquest conjunt de dades està més encaminat a l'anàlisi de les relacions d'aquest gust amb els aspectes més tècnics del cinema, que sovint són tan importants com els aspectes més artístics, però que normalment s'obliden. Dit d'una altra forma, intentarem respondre a la pregunta "com influeixen les característiques tècniques d'una pel·lícula en la seva recepció pel públic?"

Llicència

IMDb té unes condicions d'ús força estrictes i molt clares, que es poden resumir en això: "IMDb [...] grants you a limited, non-exclusive [...] license to access and make personal and non-commercial use of the IMDb Services"[4]. Així doncs, la llicència que fem servir haurà de prohibir

[1] Tohline, M. (27 gener, 2020). Tracking Mass Ideology on IMDb's Top 250: How Shifts in Societal Values Appear in the Popular Film Canon [entrada de blog]. *Bright Lights Film Journal*. Recuperat de: https://brightlightsfilm.com/tracking-mass-ideology-on-imdbs-top-250-how-shifts-in-societal-values-appear-in-the-popular-film-canon/#.YHWmpOgzZPY

[2] Lee, M. (8 octubre, 2012). IMDb Top 250 Movies List Analysis, 5th Edition [entrada de blog]. *Overthinking It*. Recuperat de: https://www.overthinkingit.com/2012/10/08/imdb-top-250-movies-5th-edition/

[3] Jain, M. (25 desembre, 2018). What I found out from IMDb top 250 movies [entrada de blog]. *Medium*. Recuperat de: https://medium.com/@mukul_jain/what-i-found-out-from-imdb-top-250-movies-1d34a68744e

[4] IMDb Conditions of Use. Recuperat de: https://www.imdb.com/conditions.

l'ús comercial d'aquestes dades. Tot i això, volem que aquestes dades es facin servir per més persones, per la qual cosa triarem una de les llicències Creative Commons de copyright públic.

Concretament, la millor llicència per a aquest cas serà la **CC BY-NC-SA 4.0**, que prohibeix l'ús comercial de les dades, obliga a fer referència a l'autor original i també obliga que les redistribucions que se'n facin (o anàlisis posteriors) hagin de distribuir-se amb la mateixa llicència. Tothom que compleixi amb aquestes restriccions sobre el dataset en podrà fer un ús lliure.

Codi

El codi font del *scraper* s'hi pot trobar a https://github.com/hectorgut47/tipologia-prac1.

Publicació del dataset

El dataset com a fitxer CSV està penjat al repositori Zenodo, concretament a l'URL: https://zenodo.org/record/4688425.