Marc Ruiz Marcos
Novembre 2021

Pràctica 1 – Tipologia i cicle de vida de les dades

A continuació es donen les respostes a les qüestions plantejades a l'enunciat de la pràctica.

Context

IMDb és una pàgina web coneguda per oferir una àmplia base de dades de l'entorn cinematogràfic, oferint així puntuacions i ressenyes, a més d'informació addicional d'actors, pel·lícules, sèries, etc.

Títol

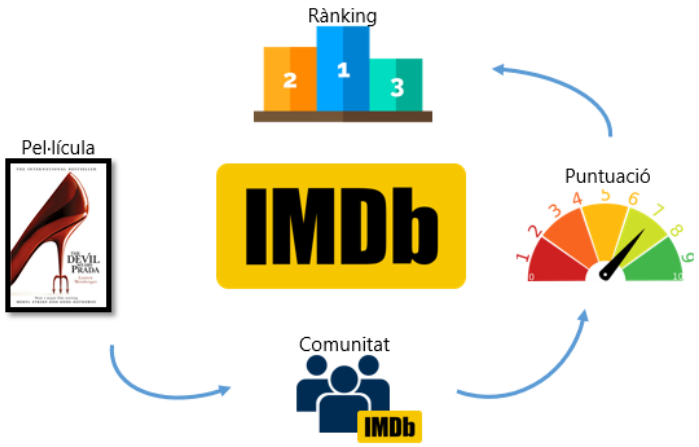
El títol escollit pel dataset és el següent: "Top 1000 de pel·lícules segons IMDb".

Descripció del dataset

Entre d'altres funcionalitats, IMDb permet als seus usuaris donar una puntuació d'una pel·lícula o sèrie de l'1 al 10. A partir d'aquesta puntuació, IMDb publica un rànking del top 1000 de pel·lícules - i és d'aquest rànking d'on es recullen les dades d'aquest dataset. Alguns dels camps inclouen la puntuació, la posició al rànking, el títol de la pel·lícula o l'any d'estrena.

Representació gràfica

A continuació es presenta un esquema explicatiu del procés a partir del qual es generen les dades recollides.



Marc Ruiz Marcos
Novembre 2021

Contingut

Els camps capturats per a cada pel·lícula són els següents:

- Rank: posició al rànking
- Title: títol de la pel·lícula
- Release year: any de llançament de la pel·lícula
- Runtime: durada en minuts de la pel·lícula
- Genre: gènere(s) de la pel·lícula, per exemple: crim i drama
- Rating: puntuació de la pel·lícula. És el valor pel qual s'ha ordenat aquest ranking.
- Director(s): director(s) de la pel·lícula
- Votes: quantitat de vots dels usuaris que s'han compatibilitzat
- Gross: guanys en milions de dòlars de cada pel·lícula en cas que es conegui

Agraïments

Com s'ha explicat anteriorment, les dades han estat extretes de IMDb, concretament d'aquesta URL. La última extracció es va fer el 26 d'octubre de 2021. Per a fer-ho, he fet servir el llenguatge de programació Python fent servir llibreries de web scraping.

Inspiració

La inspiració inicial per haver triat IMDb consisteix en què és una pàgina que faig servir bastant i que altres aplicacions (com ara PopcornTime) fan servir per a complementar la seva informació. Per tant, trobo interessant haver après l'estructura d'una de les pàgines de IMDb i l'extracció de les seves dades. El dataset obtingut ens permet fer-nos preguntes com ara:

- Hi ha correlació entre els guanys d'una pel·lícula i la seva qualitat/rànking?
- Quins directors tenen més èxit? I quins gèneres?
- Quina influència té l'antiguitat d'una pel·lícula en la seva qualitat o puntuació?
- Quina duració tenen les millors pel·lícules? És una dada prou significativa com per a extreure'n conclusions?

Llicència

Marc Ruiz Marcos
Novembre 2021

Codi

El codi font es pot trobar en aquest repositori de github.

Dataset

El dataset també es pot trobar al repositori de github de l'apartat anterior, però tal i com s'ha demanat també s'ha pujat a Zenodo. El DOI corresponent és el següent: 10.5281/zenodo.5655513 i pot ser accedit a partir d'aquest enllaç.