

Tipologia i cicle de vida de les dades

Pràctica 1: Web scraping

Alumnes:

Ana Cortes Besoli

Joan Oliva Costas

Estudis: Màster en ciència de dades

Curs: 2020-2021

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

L'objectiu d'aquesta pràctica és extreure un conjunt de dades relacionades amb pel·lícules a través de **IMDB** (Internet Movie DataBase), una pàgina web que conté tota la informació sobre contingut audiovisual (pel·lícules, sèries, documentals...) produït en tota la història.

En aquesta pràctica, es simula que som una productora audiovisual de pel·lícules que volem analitzar les dades d'altres títols, propis o no de la productora. Es desitja obtenir informació d'un conjunt de pel·lícules per tal de donar suport a la presa de decisions de la productora. Aquest suport pot anar dirigit a diversos departaments com màrqueting, producció...

En concret, la productora sol·licita informació sobre un conjunt de pel·lícules (que es defineixen a partir d'un nom i un any de llançament de forma opcional), i el programa ha d'obtenir, per cada una d'aquestes pel·lícules, la informació bàsica, així com les crítiques dels usuaris per aquelles pel·lícules. Cada vegada que es realitza una petició s'associen unes metadades a ella, per tal de poder comparar crítiques més endavant en el temps.

IMDB es centra en oferir tota la informació disponible sobre una producció audiovisual. Però disposa d'un valor afegit. Són els propis usuaris de la web els encarregats de proporcionar gran part del contingut a través de **valoracions i crítiques**. Aquesta informació té un valor molt gran per a l'anàlisi, i per tant és un dels objectius principals d'aquest projecte.

El projecte es realitza a través d'un procés de Scraping Web. En concret, s'utilitza Selenium com a llibreria, ja que la pàgina web requereix d'interacció per tal d'obtenir la part més valuosa de la informació. El llenguatge de programació escollit és Python 3.

IMBD no disposa d'una API pública. I, encara que existeixen serveis que actuen com a intermediaris per a oferir una API a partir de scraping realitzat sobre IMDB, tenen limitacions ja que són de pagament. A més, la gran majoria no ofereix la capacitat de cercar crítiques d'usuaris. Per tant, és preferible implementar un script que realitzi scraping per tal d'obtenir tota la informació que el projecte requereix sota demanda.

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

El nom que triem pel nostre Dataset és: “**Crítiques personals a pel·lícules de la plataforma IMDB**”.

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s’ha extret (es necessari que aquesta descripció tingui sentit amb el títol triat).

Les dades que s’extreuen de la pàgina web representen pel·lícules i crítiques. Es diferencia entre diverses entitats que són relacionades entre si.

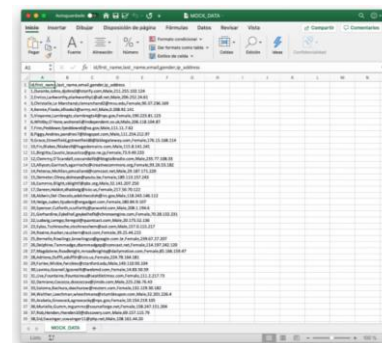
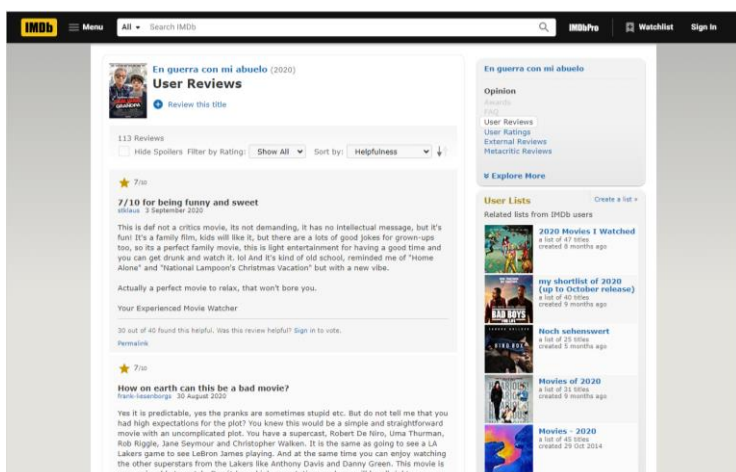
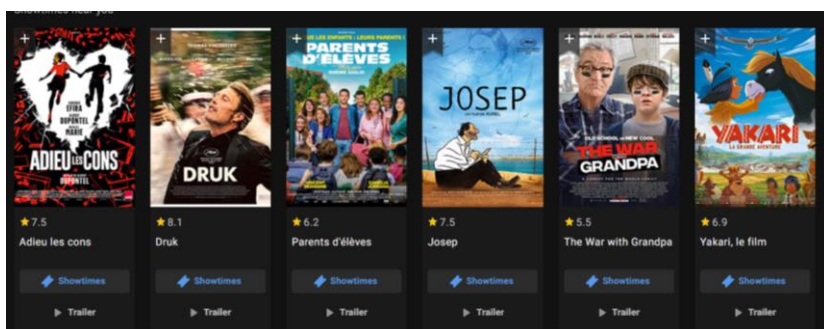
En primer lloc, la **pel·lícula**. Una pel·lícula té un identificador i les seves dades bàsiques. Cada pel·lícula a IMDB disposa d’una pàgina pròpia amb tota la informació bàsica d’aquesta.

En segon lloc, els **gèneres**. Cada pel·lícula pot tenir un o més gèneres. Els gèneres permeten cercar i classificar les pel·lícules dins de categories.

En tercer lloc, existeix l’entitat de les **crítiques**. Les crítiques són publicades pels usuaris. Cada crítica pertany a una pel·lícula. Les crítiques són una valoració per part de l’usuari sobre la pel·lícula en qüestió.

Per últim, també s’ha d’esmentar les **metadades de l’execució**. Execucions diferents poden referenciar a crítiques i pel·lícules diferents. Per tant, s’ha de considerar emmagatzemar-les com una entitat més.

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



Recollir beneficis



Produir pel·lícula



Anàlisi i presa de decisions

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

Quan s'executa el script, aquest cerca en el lloc web de IMDB en l'actualitat. Per tant, el període de temps de les dades comprén tot el passat fins al moment actual.

Les úniques dades que poden variar de forma significativa entre execucions són les crítiques. Els usuaris poden anar introduint noves crítiques al sistema.

Existeix un paràmetre d'execució del script que especifica el nombre màxim de crítiques a obtenir per cada pel·lícula. A més, també és possible especificar el criteri d'ordenació de les crítiques (entre els que ofereix IMDB). Per tant, les crítiques obtingudes en una execució poden variar també en funció d'aquests paràmetres.

Les metadades d'una execució contenen la data exacta en què s'ha realitzat l'extracció i tota la configuració de les crítiques. Per tant, les metadades contenen tota la informació per a reviu les circumstàncies d'obtenció de les crítiques.

És possible que algunes dades no estiguin informades, és a dir, siguin *missings*. Això es deu a què IMDB pot no informar d'alguns atributs (un usuari pot decidir no puntuar una crítica, per exemple). En aquest cas, el camp es deixa buit en el fitxer resultant.

La majoria de dades es processen i s'emmagatzemen amb el seu tipus intencionat (per exemple, les dates extretes sempre es guarden en format dd/MM/yyyy hh:mm:ss).

Sovint la informació requerida es situa en pàgines diferents dins del lloc web. Els enllaços concrets són semi coneguts, es a dir, se sap a quina direcció cal accedir però es desconeixen els paràmetres concrets que s'han d'enviar a la petició. Enfront d'aquest problema es requereix interaccionar amb el document web per a realitzar cerques específiques. Això es detalla a continuació.

Els resultats finals s'emmagatzemen en fitxers csv. Cada entitat té el seu propi fitxer amb una fila per cada element extret.

Com es cerquen les pel·lícules?

Els paràmetres d'entrada del programa inclouen una llista de pel·lícules. Aquesta llista conté, per cada element, el títol de la pel·lícula desitjada i l'any de llançament (ja que varies pel·lícules es poden dir igual però ser d'anys diferents).

Les pàgines de les pel·lícules es poden trobar en direccions a l'estil de <https://www.imdb.com/title/ttX/>, ón X és una cadena identificadora pròpia de IMDB.

Aquest fet suposa un problema, ja que, idealment, l'objectiu del projecte és cercar les pel·lícules a través del títol i de la data de llançament. Utilitzar el títol i l'any en comptes de l'identificador d'IMDB és millor ja que no obliga al responsable de l'execució de conèixer els identificadors a priori.

La solució aplicada per tal d'obtenir la URL d'una pel·lícula en funció del nom i de l'any de llançament consisteix en utilitzar el cercador d'IMDB.

En aquest punt apareix un nou problema, i és que IMDB no recomana l'ús de les rutes relacionades amb les capacitats de la cerca en el seu fitxer robots.txt¹.

La solució aplicada consisteix a utilitzar el buscador en l'apartat superior de la pàgina principal. Aquest buscador realitza peticions ajax segons els criteris introduïts i carrega els resultats en la mateixa pàgina.

Un cop els resultats s'han carregat, es poden recórrer amb l'ús de l'eina de scraping fins a trobar el títol més indicat (aquell que conté el títol i llançat en l'any especificat).

D'aquest procés no s'extreu cap dada final. Únicament es determina l'identificador d'IMDB de la pel·lícula. Aquest permet navegar per les pàgines que allotgen la informació de la pel·lícula.

Dataset Pel·lícula (movies)

Quan es disposa de la URL de la pel·lícula, és possible rascar les seves dades bàsiques. Per exemple, la pàgina principal de Titanic es situa en la direcció:

https://www.imdb.com/title/tt0120338/?ref=fn_tt_tt_2

Les dades del dataset de pel·lícules s'enumeren a continuació:

- **movieId**: identificador de la pel·lícula a IMDB. Per exemple *tt0120338*, És una cadena de text.
- **movieTitle**: títol de la pel·lícula. Si la pel·lícula té títol original (ja que IMBD mostra el títol de la pel·lícula de la cultura per defecte del navegador), s'utilitza el títol original en comptes del títol cultural. Cadena de text.
- **movieYear**: any de llançament de la pel·lícula. Cadena de text.
- **Duration**: duració de la pel·lícula en minuts. És un número sencer (minuts de duració)
- **Rating**: mitjana puntuació dels usuaris. És un número decimal.
- **nVotes**: nombre de vots totals dels usuaris. Número Sencer.
- **Popularity**: rànquing de popularitat de la pel·lícula. Pot variar entre execucions. Número Sencer.
- **BudgetUSD**: pressupost de la pel·lícula. De tipus monetari.

¹ Les rutes /find i /search no són recomanades pel robots.txt. Aquesta informació es troba a l'annex.

- **GrossUSD**: recaptació bruta de la pel·lícula arreu del món. De tipus monetari.

Algunes dades poden no aparèixer en la pàgina de la pel·lícula (p.ex. La popularitat). En aquest cas la dada es queda en blanc en els fitxers de sortida.

Les dades s'extreuen a través de l'examinació del document HTML. Els camps enumerats apareixen sempre dins d'etiquetes amb una classe concreta, pel que són fàcilment localitzables. Un cop s'ha extret el contingut de l'etiqueta (text), s'apliquen transformacions per obtenir el resultat desitjat.

Per exemple, el text del pressupost apareix a IMDB com a: "Budget: \$400,000,000", pel que és necessari extreure únicament la quantitat. En la majoria de vegades s'utilitzen expressions regulars per a processar el text.

Dataset Gèneres (genres)

Els gèneres de les pel·lícules s'obtenen de la mateixa URL que la pel·lícula. Al ser un conjunt arbitrari, es decideix aplicar un nivell de normalització i separar ambdues entitats. Un gènere conté:

- **movieId**: identificador de la pel·lícula. Cadena de text.
- **genre**: Cadena de text (p.ex. drama, romance, action, ...etc).

Dataset Crítiques (reviews)

Les crítiques d'una pel·lícula es poden llegir en la seva totalitat a través d'una pàgina individual. Per exemple, les crítiques dels usuaris sobre la pel·lícula Titanic es troben a https://www.imdb.com/title/tt0120338/reviews?ref=tt_ov_rt.

Per tant, també és indispensable disposar de l'identificador de la pel·lícula per obtenir la seva llista de crítiques. D'aquesta entitat, interessa obtenir totes les dades possibles. Es poden arribar a extreure les següents:

- **MovieId**: identificador de la pel·lícula que referencia la crítica.
- **Username**: nom de l'usuari que realitza la crítica. Cadena de text.
- **Date**: data de la crítica. No conté el temps exacte, sols el dia. És un datetime en format dd/mm/yyyy.
- **Review_title**: títol de la crítica. Cadena de text.
- **Rating**: puntuació que la crítica atorga a la pel·lícula. Va del 0 al 10. És un número sencer.
- **Text**: comentari de la crítica. És una cadena de text.
- **helpfulYes**: nombre d'usuaris que han trobat útil la crítica. Número sencer.

- **helpfulTotal**: nombre d'usuaris que han votat sobre la utilitat de la crítica. Número sencer.
- **isSpoiler**: indica si la crítica conté spoilers o no. És un booleà.

En aquest cas, la informació no es troba de forma unitària a la pàgina, sinó que és repetida. Cada crítica es troba dins del seu divisor HTML. Per tant, la forma de llegir les crítiques consisteix en recórrer tots els contenidors de crítiques i extreure les dades per cada una d'elles.

Existeixen problemàtiques associades a aquest algorisme. En primer lloc, el nombre de crítiques que apareix en la pàgina per defecte és limitat (al voltant de 25). Això ho realitza IMDB per tal de no carregar totes les crítiques alhora. Les crítiques són paginades. Obrir una nova pàgina requereix fer clic a un botó al peu de la pàgina.

Aquest problema es pot solucionar amb l'ús de **Selenium**. Es pot programar un algorisme que faci clics sobre el botó de càrrega tantes vegades com sigui necessari, esperant que la pàgina es refresqui entre clic i clic.

Per motius de rendiment, es pot especificar, en l'execució del programa, un paràmetre que determina quin nombre màxim de crítiques es vol obtenir. Aquest paràmetre serà **max_reviews**.

En segon lloc, la llista de crítiques apareix ordenada per un criteri arbitrari. La pròpia pàgina ofereix ordenar les crítiques en funció d'alguns criteris com la utilitat, la rellevància, la puntuació, etc. Això entra en conflicte amb el fet anterior. Per tant, s'ha de considerar dos paràmetres extra en l'execució. Aquests paràmetres són **sort_by**, que determina el criteri d'ordenació de la llista de crítiques, i **sort_order**, que indica l'ordre (ascendent o descendent).

Materialitzar aquests dos últims paràmetres resulta trivial ja que es poden incloure a la URL de la pàgina de les crítiques. Per a les del Titanic, per exemple, resultaria en la següent URL: <https://www.imdb.com/title/tt0120338/reviews?sort=totalVotes&dir=desc>

Dataset metadades de l'execució (metadata)

Les metadades de l'execució permeten identificar el període d'obtenció de les dades i el context d'execució detallat. Aquestes són:

- **date**: data d'execució del script. És en format dd/MM/yyyy hh:mm:ss
- **sort_by**: criteri d'ordenació de les crítiques. Els tipus disponibles es poden consultar a la documentació del repositori.
- **sort_order**: asc | desc segons ordenació preferida de les crítiques.
- **max_reviews**: nombre màxim de crítiques a obtenir per cada pel·lícula

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

El procés de recerca ha consistit en determinar, en primera instància, si es podia fer ús d'una API, interna o externa, per tal d'obtenir la informació proposada.

En primer lloc, s'ha realitzat recerca sobre com recopilar informació de la web de IMDb sense necessitat de realitzar scraping. S'ha determinat que **IMDb té una web de desenvolupadors**, (<https://developer.imdb.com/>).

En aquesta web s'explica que existeix una API i tres productes d'extracció massiva de dades que utilitzen la tecnologia "AWS Data Exchange". Aquest és un servei de càrrega i anàlisi de dades al núvol ofert per Amazon.

El procediment per a consumir les APIs no es detalla. Tot i així, sí que s'especifica que per obtenir més informació, et posis en contacte amb l'equip d'IMDb a través d'una direcció de correu facilitada.

Així doncs, s'ha redactat un correu demanant informació sobre l'ús de les API's i els seus sistemes, evidenciant el context de que es tracta d'una pràctica acadèmica sense cap intenció comercial. La resposta es pot resumir en els següents punts:

- Les seves API's i serveis són de pagament.
- L'ús de la informació que es pugui extreure d'IMDb no pot ser utilitzada amb propòsits comercials.
- Si es volen utilitzar les dades amb finalitats acadèmiques, es pot fer sempre i quan es citi l'autor original com a IMBD.
- Les API's externes independents de IMDb no són reconegudes per la companyia i per tant no s'haurien d'utilitzar.
- Recomanen l'ús de datasets públics ja processats i publicats en la seva pròpia web.²

Addicionalment, també s'ha cercat per internet aquestes aplicacions externes que ofereixen una API que treballa sobre IMBD. S'han trobat varis exemples com [rapidapi](#) o [omdbapi](#). Encara que no s'han utilitzat per diversos motius. En primer lloc, perquè la majoria són de pagament. En segon lloc, perquè el propi equip d'IMDb no recomana el seu ús. I en tercer lloc, perquè no oferien la possibilitat d'extreure les crítiques dels usuaris.

També s'han trobat exemples d'aplicacions software que utilitzen algunes de les API's externes per a obtenir directament les dades de les pel·lícules, sense necessitat de treballar amb una API.³

² La URL d'accés és: <https://www.imdb.com/interfaces/>

³ Es pot trobar el software a <https://www.neoteo.com/como-crear-un-catalogo-de-peliculas-con-datos-de-imdb/>

Per últim, s'ha realitzat recerca sobre altres articles relacionats amb l'extracció de dades d'IMDB per tal d'ampliar el camp de visió sobre la realització de la pràctica. La majoria de casos no fan ús de cap interacció, sinó que es limiten a analitzar pàgines de forma estàtica utilitzant llibreries com BeautifulSoup. Cap d'elles extreu les crítiques d'un títol.⁴

Per tant, es pot realitzar la pràctica sempre i quan no es vulnerin els requeriments que exigeix la companyia. Es descarta utilitzar les dades públiques que recomanen, ja que no permeten dur a terme l'objectiu de la pràctica (scraping), i, a més, tampoc no inclouen les crítiques dels usuaris.

Respecte als **agraïments** i al **propietari** de les dades de la pràctica,

En primer lloc, es reconeix a **IMDB** com al propietari de les dades que es puguin extreure en el context d'aquesta pràctica i s'assegura que aquestes sols s'utilitzen amb finalitats acadèmiques i no comercials. També s'agraeix la seva cortesia a l'hora de permetre dur a terme la pràctica.

Addicionalment, també s'agraeix de forma especial al Sr. Riley per a la seva ràpida resposta i explicació sobre el tema.

Cal dir que, si aquest estudi es realitzés sota condicions reals, s'hagués intentat arribar a un acord comercial per a utilitzar les dades necessàries per a dur a terme el projecte.

En **segon lloc**, s'agraeix a **tots els usuaris** que han contribuït amb les seves crítiques a la base de dades d'IMDB. Gràcies a ells es pot obtenir la informació necessària sobre les crítiques per analitzar-la, extreure coneixement i aplicar decisions en la situació simulada per aquesta pràctica.

⁴ Els articles es poden consultar a la bibliografia.

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

L'escenari proposat és la simulació d'un estudi d'una productora de cinema. Aquesta considera que amb l'ús de les noves tecnologies es pot obtenir informació molt útil a l'hora de produir noves pel·lícules.

La productora és conscient que l'opinió del públic és un dels factors claus a l'hora de determinar l'èxit d'una pel·lícula. Amb l'ajuda d'IMDB i de la seva base de dades, es pot extreure informació que els consumidors aporten en forma de crítiques. Conèixer l'opinió del públic pot ser crucial per maximitzar els beneficis econòmics o culturals.

Una de les tècniques que es poden aplicar al dataset resultat és l'anàlisi de sentiment del text. A partir de totes les dades de les crítiques, i en especial dels comentaris, es pot extreure una quantitat d'informació molt valuosa que reflexa l'opinió general que té el públic sobre una pel·lícula.

Aquesta informació es pot traduir en coneixement i en una presa de decisions que permeti millorar la qualitat de futures pel·lícules. Per exemple, pot significar conèixer quins gèneres cal potenciar, quins actors són els més preferits pel públic...

Algunes de les preguntes que es pretén respondre són:

- Quin és el gènere que els usuaris millor valoren?
- Quins gèneres de pel·lícules de la productora són els pitjors i millors valorats?
- Quins són els actors/actrius que més agraden. Inclús es pot segmentar per gènere. Fent anàlisi de sentiment de les crítiques podrem trobar opinions positives o negatives sobre actors i/o pel·lícules.⁵
- Quines pel·lícules són les que han tret més benefici en funció de la seva opinió?
- Quines pel·lícules valoren més els usuaris?
- En quines èpoques de l'any tenen més èxit les pel·lícules?
- Moltes altres...

Per tal d'acotar l'extracció de dades de l'apartat 10 sota un escenari concret, es decideix aplicar el següent cas d'ús:

Som la productora responsable de la saga de El Padrino (I, II i III). Es vol realitzar un estudi sobre quina és l'opinió del públic de les tres pel·lícules, per tal d'estudiar si resulta viable realitzar una quarta entrega. A partir de la informació obtinguda, es podrà determinar si la quarta entrega és esperada o no, i quins elements d'entregues anteriors s'haurien de potenciar

⁵ Encara que no s'extreuen els actors de les pel·lícules (en sortir de l'abast de la pràctica). Es poden localitzar els actors a partir dels comentaris de les crítiques.

i/o eliminar. Es desitja obtenir les 1000 primeres crítiques de cada pel·lícula ordenades per utilitat.

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

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

La llicència que més encaixa en el projecte és: **Released Under CC BY-NC-SA 4.0 License**

Aquesta llicència aplica bastantes restriccions. S'ha escollit pels següents motius:

En primer lloc, no permet l'ús de la informació amb propòsits comercials, però sí que en permet la compartició i l'adaptació. Aquestes clàusules concorden amb les restriccions que l'equip d'IMDB ha imposat sobre l'ús acadèmic de les dades que ha cedit.

En segon lloc, la llicència exigeix l'atribució de l'autor original de les dades, que també és consonant amb les demandes del mateix equip. Com ja s'ha comentat anteriorment, s'atribueix la propietat de les dades a IMDB i s'agraeix la seva cortesia.

Per últim, la llicència permet adaptar i transformar les dades sempre i quan aquestes es publiquin sota la mateixa llicència original. Les possibles publicacions de les dades processades actuaran sota aquesta llicència.

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

La URL del repositori és: <https://github.com/Jolicost/WebScraping>

10. Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

La URL de Zenodo és: <https://zenodo.org/record/4155443#.X5rtte2CHcc>

Taula de contribucions

Contribucions	Signa
Recerca prèvia	ACB, JOC
Redacció de respostes	ACB, JOC
Desenvolupament codi	ACB, JOC

Bibliografia

Akdag, M. (2020). *Web Scraping from IMDB Site*. [data consulta: 27 d'octubre de 2020]. Disponible a: <https://www.kaggle.com/akdagmelih/web-scraper-imdb-movies>

AWS Data Exchange. [data consulta: 27 d'octubre de 2020]. Disponible a: <https://aws.amazon.com/es/data-exchange/>

Benabaya, I. (2020). *How to Scrape TV Show Episode IMDb Ratings Using Python & BeautifulSoup* [data consulta: 27 d'octubre de 2020]. Disponible a: <https://towardsdatascience.com/scraping-tv-show-episode-imdb-ratings-using-python-beautifulsoup-7a9e09c4fbe5>

Campbell, J. (2020). *Beginner's Guid on Web Scraping in R (using rvest) with hands-on exemple*. [data consulta: 27 d'octubre de 2020]. Disponible a: <https://rpubs.com/uky994/578161>

Fernández, E. (2018). *Cómo crear un catálogo de pel·lícules con datos de IMDb* [data consulta: 27 d'octubre de 2020]. Disponible a: <https://www.neoteo.com/como-crear-un-catalogo-de-peliculas-con-datos-de-imdb/>

IMDb: [data consulta: 12 d'octubre de 2020]. Disponible a: <https://www.imdb.com/>

IMDb developer: [data consulta: 21 d'octubre de 2020]. Disponible a: <https://developer.imdb.com/>

Licencias Creative Commons [data consulta: 27 d'octubre de 2020]. Disponible a: https://es.wikipedia.org/wiki/Licencias_Creative_Commons

Predum, R. (2019). *Web scraping Sci-Fi movies from IMDB with Python* [data consulta: 27 d'octubre de 2020]. Disponible a: <https://medium.com/@rileypredum/web-scraping-sci-fi-movies-from-imdb-with-python-48fa3f85723c>

Quora. (2019). *Is it legal to scrape imdb and store basic data on each item?* [data consulta: 27 d'octubre de 2020]. Disponible a: <https://www.quora.com/Is-it-legal-to-scrape-imdb-and-store-basic-data-on-each-item>

Sobre las licencias [data consulta: 27 d'octubre de 2020]. Disponible a: <https://creativecommons.org/licenses/?lang=es>

¿Qué tipo de licencias deben acompañar a los datasets en acceso abierto? [data consulta: 27 d'octubre de 2020]. Disponible a: <http://bibliotecas.csic.es/es/node/300>

Annex

Fitxer robots.txt de la web <https://www.imdb.com/>

```
# robots.txt for https://www.imdb.com properties
User-agent: *
Disallow: /OnThisDay
Disallow: /ads/
Disallow: /ap/
Disallow: /mymovies/
Disallow: /r/
Disallow: /register
Disallow: /registration/
Disallow: /search/name-text
Disallow: /search/title-text
Disallow: /find
Disallow: /find$
Disallow: /find/
Disallow: /tvschedule
Disallow: /updates
Disallow: /watch/_ajax/option
Disallow: /_json/video/mon
Disallow: /_json/getAdsForMediaViewer/
Disallow: /list/ls*/_ajax
Disallow: /*/*/*rg*/mediaviewer/rm*/tr
Disallow: /*/*/*rg*/mediaviewer/rm*/tr
Disallow: /*/*mediaviewer/*/*tr
Disallow: /title/tt*/mediaviewer/rm*/tr
Disallow: /name/nm*/mediaviewer/rm*/tr
Disallow: /gallery/rg*/mediaviewer/rm*/tr
Disallow: /tr/
Disallow: /title/tt*/watchoptions
```