TIPOLOGIA I CICLE DE VIDA DE LES DADES. PRÀCTICA 1. WEB SCRAPING

AUTORS (PER ORDRE ALFABÈTIC)

Joan Ginard Illescas

Miquel Piña Grau

Contenido

1. CONTEXT

Per aquesta pràctica vam decidir centrar-nos en pàgines la temàtica de les quals ens fes fàcil imaginar-nos el possible joc de dades final i la presentació d'aquest com a CSV.

Per aquest motiu ens vam fixar en dos tipus de pàgines: les immobiliàries i les de cinema. Vam descartar les primeres perquè algunes dades podria no ser convenient o adequat publicar-les a un CSV i per això ens vam decantar per les segones.

De totes elles vam trobar la pàgina de CINEMANIA que ofereix una constant actualització de pel·lícules a mesura que són estrenades i que afegeix una valoració de la pel·lícula, es a dir, obteníem la fitxa tècnica de la pel·lícula juntament amb una puntuació o nota, de tal manera que al joc de dades final podríem filtrar per país, puntuació, actors...

No obstant ens vam trobar amb un problema i és que abans de juny del 2014 les pel·lícules no tenien les dades en una fitxa tècnica i la forma de proporcionar les dades podia variar d'una pel·lícula a l'altra. Per tant hi ha un límit de data per a les pel·lícules de les quals podem recuperar la fitxa

2. TÍTOL DATASET

Darreres estrenes cinematogràfiques. Des d'avui fins a juny del 2014.

3. DESCRIPCIÓ DEL DATASET

El conjunt de dades són 2400 pel·lícules presentades des de les estrenades la darrera setmana[1] fins a juny de 2014. Tot i que en l'execució de l'scraper podem triar una data final anterior i recollir menys pel·lícules.

Juntament amb els títols es recullen tota una sèrie de dades típiques: gènere, país de procedència, director, guionista... (veure punt 5)

Les dades no han estat netejades o pre-processades i per tant pot haver dades amb formats inconsistents, per exemple normalment la durada té el format de número + "min." però en algun cas ens trobem només un número o també "minutos" enlloc de "min".

També trobem alguna errada en la introducció de les dades per part de la pàgina, i per exemple ens trobem la durada de la pel·lícula a l'edat recomanada.

4. REPRESENTACIÓ GRÀFICA

[1] Execució de l'script per obtenir les dades : 24 – 10 – 2020 .

5. CONTINGUT

Les dades de les pel·lícules que figuren a la pàgina es remunten a abans de l'any 2010, però com s'ha comentat no es comencen a presentar en un format de fitxa tècnica estandarditzat fins a juny del 2014.

Els camps que inclou el nostre data set són els següents:

- **Índex**: un número enter que fa d'índex de la pel·lícula per facilitar el recompte dels registres en el moment de recollir les dades
- **Títol:** El títol de la pel·lícula
- **Valoració:** Un número enter entre 1 i 10 que ens dona una "nota" de la pel·lícula segons el crític
- **Data_crítica:** Data de la crítica
- **Gènere:** Gènere de la pel·lícula, pot incloure més d'un. Per exemple: "Comedia, Romance"
- **Director:** Nom de la persona o persones que han dirigit la pel·lícula
- **Intèrprets:** Nom de la persona o persones que han protagonitzat la pel·lícula
- **Guio:** Nom de la persona o persones que han escrit la pel·lícula
- **País:** Nom del país de producció de la pel·lícula
- **Durada:** Durada de la pel·lícula en minuts.
- **Edat_recomanada:** Edat mínima recomanada per poder veure la pel·lícula
- **Distribuidora:** Nom de l'empresa distribuïdora de la pel·lícula
- **Data_estrena:** Data d'estrena de la pel·lícula. És important fer notar que sempre coincideix amb la data de la crítica fins ara.

Voldríem indicar que els camps poden estar buits, no només perquè hi ha pel·lícules on alguns camps podrien no tenir sentit com el intèrprets a un documental (a alguns sí) sinó també perquè la fitxa no conté dades dels mateixos.

Per recollir les dades vam fer servir un web_scraper en llenguatge Python (3.7) i fent servir la llibreria BeautifulSoup.

Per poder fer el web scraping d'aquesta pàgina havíem de recórrer la pàgina, recollir el títol i valoració de la pel·lícula i entrar al link de la pel·lícula ja que només aquí es troba la fitxa tècnica de la pel·lícula. Finalment havíem de passar a la següent pàgina i repetir el procés, així fins a obtenir les dades de totes les pel·lícules.

6. AGRAÏMENTS

Volem agrair al lloc web que faciliti el web scraping sense posar cap limitació o gairebé cap limitació (robots.txt no impedeix cap lloc dins el website)

També hem d'agrair la inspiració a l'hora d'incloure arguments en l'script aportada pel treball previ que hem trobat en aquest github i inclòs en l'enunciat:

- https://github.com/rafoelhonrado/foodPriceScraper

7. INSPIRACIÓ

El conjunt de dades pot ser interessant per fer un anàlisi des de diferents caires del tipus de pel·lícules que s'estrenen

En concret podrem contestar preguntes sobre la relació hi ha entre la valoració i la resta de camps. Per exemple: Les pel·lícules més llargues tenen tendència a tenir més valoració? Quina distribuïdora té les millors pel·lícules?

També podrem contestar preguntes sobre l'evolució de les pel·lícules al llarg del temps: Van canviant el gèneres predominants amb els anys? Quins gèneres es posen de moda? Quin intèrprets pugen a l'estrellat o protagonitzen més pel·lícules darrerament?

Finalment podem trobar altres relacions: Hi ha països on es produeixin principalment pel·lícules d'un determinat gènere? Hi ha països amb tendència a produir pel·lícules més llargues?

8. LLICÈNCIA

Primer examinem els diferents tipus de llicències:

- **Llicència CC0**: En aquest cas la l'obra és entregada al domini públic renunciant a tots els drets de propietat intel·lectual sobre ella. Per tant, l'obra pot ser copiada, modificada, distribuïda fins i tot amb finalitat comercial. Però d'altra banda l'autor de l'obra no ofereix cap garantia sobre ella ni es fa responsable del seu ús posterior.
  Això no eximeix que patents o propietats intel·lectuals inclosos en la obra no quedin protegits, es a dir, tot i que l'obra tingui aquesta llicència si inclou algun tipus de propietat intel·lectual aquesta segueix estant protegida.
- **Llicència CC BY-NC-SA 4.0:** Les obres sota aquesta llicència es poden copiar i redistribuir i a més es poden modificar però sempre mantenint 3 condicions:
  - **Reconèixer l'autoria** i si és el cas, indicant els canvis realitzats sobre l'original. En cap cas es pot donar peu a fer pensar que l'autor original dona suport o patrocina l'ús que en fem
  - **No comercial**
  - **Redistribuir amb la mateixa llicència**, en cas de fer modificacions la nova obra s'ha de distribuir sota la mateixa llicència
- **Llicència CC BY-SA 4.0:** És semblant a l'anterior només que ara SÍ es **permet la finalitat comercial**
- **Llicència ODBL** (Open Database License): Fa referència a una base de dades, en aquest cas hi ha llibertat per compartir modificar i utilitzar la base de dades. No obstant les dades contingudes a la base de dades (imatges, material audiovisual...) segueixen protegides per la llicència que pertoqui. Es a dir, una base de dades sota aquesta llicència pot tenir diverses llicències per diversos continguts.

Nosaltres finalment hem triat la **llicència CC BY-SA 4.0**, doncs considerem que les dades de les pel·lícules són públiques i possiblement no hi ha problema en aquesta reutilització, tot i que a la descripció de zenodo indiquem que el treball no té intenció comercial

9. CODI

El codi es troba al repositori creat de github i compartit pels dos autors del projecte tot i que aquí figuri el link des de l'usuari d'un d'ells.

https://github.com/JoanGinard/cinema_scraper

10. DATASET

El data set es troba tant al repositori de github com a zenodo on té el DOI següent:

```
10.5281/zenodo.4126260
```

REALITZACIÓ

CONTRIBUCIONS	AUTORS
Recerca prèvia	Joan Ginard i Miquel Piña
Redacció de les respostes	Joan Ginard i Miquel Piña
Desenvolupament codi	Joan Ginard i Miquel Piña

BIBLIOGRAFIA

(No s'inclou però també s'ha consultat la bibliografia recomanada a l'enunciat de la pràctica)

CREATIVE COMMONS. *CC0 1.0 Universal (CC0 1.0) Oferiment al Domini Públic*. [en línia] [Data de consulta: 17 d'octubre de 2020]. Disponible a: https://creativecommons.org/publicdomain/zero/1.0/deed.ca

CREATIVE COMMONS. *Reconeixement-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)*. [en línia] [Data de consulta: 17 d'octubre de 2020]. Disponible a: https://creativecommons.org/licenses/by-nc-sa/4.0/deed.ca

CREATIVE COMMONS. *Reconeixement- CompartirIgual 4.0 Internacional (CC BY-SA 4.0)*. [en línia] [Data de consulta: 17 d'octubre de 2020]. Disponible a: https://creativecommons.org/licenses/by-sa/4.0/deed.ca

OPEN DATA COMMONS. *Open Data Commons Open Database License (ODbL) v1.0*. [en línia] [Data de consulta: 17 d'octubre de 2020]. Disponible a: https://opendatacommons.org/licenses/odbl/1-0/