Sergi Mas Rodriguez – 47793952K

Tipologia i cicle de vida de les dades - PRAC 1

Consideracions prèvies:

- El motiu pel qual he decidit fer la pràctica individualment enlloc per parelles tal i com es demanava des del professorat de l'assignatura és la meva reduïda i volàtil disponibilitat horària. Aprofito estones en què la feina no m'exigeix tant o bé les estones en què els dos fills petits de casa estan dormint (per posar un exemple, aquest document l'estic escrivint a les 6h del matí d'un dissabte) o entretinguts mirant la TV. El fet de tenir un horari dependent de persones caòtiques com els meus fills em va portar a decidir que, per tal de no dependre dels desenvolupaments d'altres persones i poder treballar al meu ritme, realitzar aquesta pràctica sol.
- En aquest document intentaré anar donant resposta als diferents apartats que es plantegen a la secció de *Descripció de la pràctica a realitzar*. També s'ha realitzat un *README.md* a GitHub on es detallen les diferents parts del repositori però he cregut convenient realitzar una entrega més formal amb aquest document i una altre més pràctica amb la creació del *README.md*.

Així doncs, comencem a detallar els diferents apartats que se'ns demana per a la realització d'aquesta pràctica:

1. **Context:**

   S'ha optat per l'extracció de dades ubicades en el servidor de Filmaffinity perquè sóc usuari d'aquesta web des de fa molts anys i fent una mica de cerca per Google vaig trobar algun repositori a GitHub que s'havia dedicat a fer-ho i em va semblar molt interessant poder-ho desenvolupar també jo i veure si seria capaç o no de desenvolupar-ho (en els propers apartats posaré l'enllaç del repositori mencionat).

   Filmaffinity és un agregador de dades sobre pel·lícules, sèries, curts i documentals a l'estil de l'americana IMDB. En la fitxa d'una pel·lícula podem trobar-hi informació sobre la data de publicació, director, gènere, guionistes, repartiment, ... així com una votació dels usuaris sobre la pel·lícula en qüestió.

   Per altra banda, el que fa realment interessant Filmaffinity és el sistema de "Ànimes Bessones" on, en funció de les teves votacions, busquen altres usuaris que comparteixin afinitats cinèfiles i, a partir d'aquí, et proposen recomanacions de pel·lícules que no has vist i que les teves ànimes bessones sí. Imagino que la manera de detectar les ànimes bessones deu estar basat en un mètode de clustering i, per ser sincers, és un dels enigmes que més em desperten la curiositat.

   Per acabar aquesta primera secció, deixo alguns enllaços per ampliar la informació sobre Filmaffinity i els seus creadors:

   https://www.filmaffinity.com/es/site-guide.php

https://es.wikipedia.org/wiki/FilmAffinity

https://www.elespanol.com/economia/empresas/20170623/225978152_0.html

2. **Definir títol del data set:**

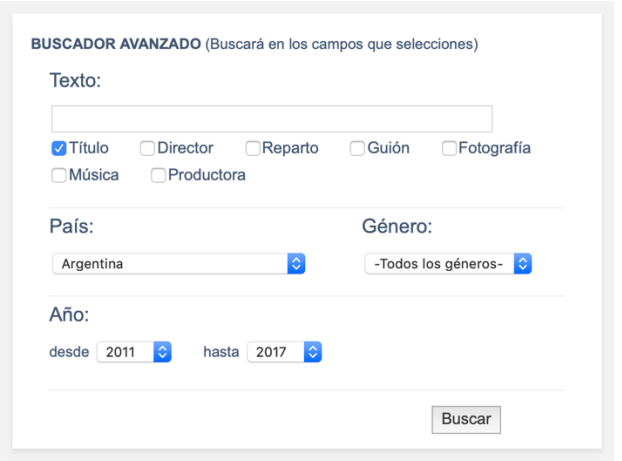
   El nom del data set és *db_FA.csv.*

3. **Descripció del data set (inclou representació gràfica i continguts):**

   El data set es basa en l'extracció de dades de pel·lícules filtrades segons any i país. Per fer-ho, ens hem basat en el motor de cerca intern que té la pròpia pàgina (https://www.filmaffinity.com/es/advsearch.php) però hem reduït, per simplicitat del programa, els països que permet seleccionar i tan sols permet seleccionar els anys. Si es volgués, es podria reproduir amb facilitat el motor de cerca que tenen implementat simplement fixant-nos en com tenen parametritzat els altres camps (gèneres, sèries de TV i documentals).

   Haig de reconèixer que la meva idea inicial era utilitzar una pàgina que han descatalogat i que permetia poder filtrar per any totes les pel·lícules però no he estat capaç de trobar-la i, en el seu defecte, he trobat aquest motor. Així doncs, he optat per utilitzar el mecanisme de funcionament d'aquest motor (ja que és molt fàcil i intuïtiu) i, a partir d'aquí, fer la cerca per cada una de les pel·lícules.

   Abans de descriure pròpiament el data set, descriuré breument el motor de cerca que utilitza FA per realitzar les cerques dins el seu repositori. Si, per exemple, volem consultar les pel·lícules rodades a Argentina. Que s'han rodat entre els anys 2011 i 2017 faríem la següent consulta mitjançant el formulari:



   Però des la URL es traduiria amb el següent enllaç:

   https://www.filmaffinity.com/es/advsearch.php?stext=&stype%5B%5D=title&country=AR&genre=&fromyear=2011&toyear=2017

   Així, com es pot veure en l'enllaç anterior, modificant alguns paràmetres de la URL és extremadament senzill realitzar el filtre i, per tant, en el nostre programa ens hem basat en aquesta simplicitat per poder explotar la seva base ja que, cada títol de

pel·lícula, té associada una ID. Per exemple, la pel·lícula argentina *Relatos Salvajes,* es pot consultar mitjançant el següent enllaç: https://www.filmaffinity.com/es/film809035.html. Així, veiem que cada pel·lícula té associada una ID única i és amb això que, primer ens guardarem una array amb totes les IDs de cada pel·lícula i llavors, simplement anar recorrent aquest array, realitzarem la cerca per cada una de les seves pàgines webs associades.

Un cop vist com funciona el seu motor de cerca i com es basa el sistema de classificació de les pel·lícules, anem a descriure pròpiament el data set. Aquest constarà dels següents atributs:

- Id
- Títol
- Any
- Duració
- País
- Director/s
- Guionista/es
- Música
- Fotografia
- Productora
- Repartiment
- Gènere
- Nota
- Votacions
- Pàgina web oficial

Cal destacar que aquestes dades no són fixes. Sense anar més lluny, la nota i el nombre de votacions varien constantment amb lo que un valor descarregat avui no té perquè ser el mateix que un descarregat demà. Addicionalment, hi ha pel·lícules antigues que no s'han anat entrant a mesura que la gent n'informava de la seva existència. Per exemple, jo personalment vaig informar de dos telefilms realitzats per TV3 i que no estaven dins la base de FA i que, gràcies al meu avís, van incorporar.

4. Agraïments i Inspiracions:

Abans que res, m'agradaria agrair a Filmaffinity la seva facilitat d'extracció de dades així com la seva existència ja que m'ha fet passar molts bons moments (gràcies a aquesta web he descobert grans joies ocultes).
Addicionalment, m'he llegit amb profunditat i he intentat entendre lo que millor he pogut el següent projecte:
https://github.com/sergiormb/python_filmaffinity
Com es pot veure, es tracta d'un web-scraping molt més avançat que el meu sobre la web de Filmaffinity i l'he utilitzat com a referència per realitzar el meu programa.
Addicionalment, també aprofito per passar l'enllaç d'un vídeo – tutorial web sobre web-scraping que he consultat per a la realització d'aquest treball. Com he dit amb

anterioritat, el meu coneixement sobre el tema abans de la realització d'aquesta pràctica era pràcticament nul: https://www.youtube.com/watch?v=NCfmEcyqgao

5. **Llicència:**

   Bastant-nos en en treball de l'usuari *sergiormb*, la seva llicència és basa en la MIT. Aquest tipus de llicència no té restriccions i permet el seu ús, còpia, modificació, integració amb altres softwares, així com publicació, distribució i ús comercial del programa. Aquests drets, però, estan subjectes que es publiqui a la pertinent publicació del copyright. En cas de no complir-se aquesta condició, no es pot fer cap ús del seu software.

   El nostre data set obtingut s'emmarca sota la llicència de *creative commons by-nc-sa 4.0*. Aquesta llicència permet copiar i redistribuir el material en qualsevol mitjà o format i adaptar-lo o modificar-lo sota les condicions de: atorgar el reconeixement apropiat al propietari de les dades i indicar si s'han realitzat cap tipus de modificacions en les dades, així com la no – utilització per finalitats comercials de les dades i que els projectes en els quals es treballa estiguin amparats sota la mateixa llicència. Per a més informació, recomano la lectura de la seva pàgina web: https://creativecommons.org/licenses/by-nc-sa/4.0/

6. **Codi i data set:**

   El codi i data set es troba al repositori de GitHub expressament creat per a la realització d'aquesta pràctica. En la wiki del repositori (https://github.com/Sergi-MR/prac1_tipologia_UOC) es detalla una breu descripció de tota la documentació penjada així com un breu resum d'aquesta memòria.