TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PEC 1

Rafael Corvillo Alonso
Pablo López Ladrón de Guevara

Máster Universitario en Ciencia de Datos

Práctica 1 – Web scraping

Contenido

1. Contexto

Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Filmaffinity es una página web española dedicada a recopilar información sobre contenido audiovisual. En su base de datos se puede encontrar las fichas completas de películas, series y documentales. Para esta práctica nos hemos centrado únicamente en la recolección de información sobre películas. Cada película recibe una puntuación calculada como la media de las puntuaciones recibidas por los usuarios, la cual también añadiremos al dataset.

IMDb es una base de datos en línea de contenido audiovisual similar a Filmaffinity. Actualmente es propiedad de Amazon. Tiene un contenido más extenso, ya que es posible encontrar información sobre videojuegos e información detallada sobre los actores y personajes ficticios de las producciones. Para este proyecto solamente nos interesa obtener la valoración media de las películas y el número de votos recibido.

Ambas páginas han sido elegidas por contener una extensa base datos sobre películas. Se puede decir que Filmaffinity es el referente hispanohablante a la hora de obtener este tipo de información e IMDb es la base de datos internacional por excelencia. Uno de los posibles usos del conjunto de datos, como se comentará más adelante, es realizar un análisis de valoración internacional frente a valoración en la comunidad hispanohablante.

2. Definir un título para el dataset

Elegir un título que sea descriptivo.

Información sobre las mejores películas de Filmaffinity y comparación con IMDb.

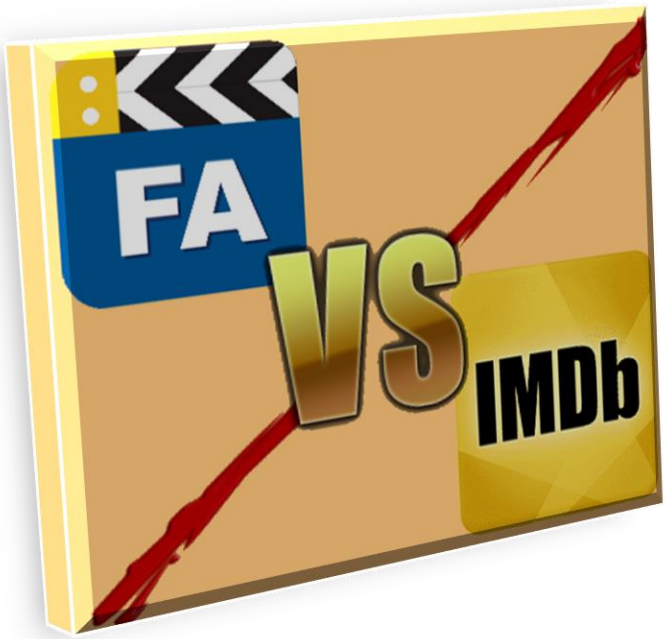
3. Descripción del dataset

Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El conjunto de datos recopila información variada sobre las mejores películas del sitio web Filmaffinity. Los campos recogidos se detallan en el punto 5 de este documento. Los dos últimos campos del dataset corresponden a la valoración y número de votos de IMDb. Dichos campos pueden ser comparados con los obtenidos de Filmaffinity. El juego de datos es almacenado como archivo CSV.

4. Representación gráfica

Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. Contenido

Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Este dataset contiene información del top de películas con mejor puntuación en Filmaffinity e incluye los siguientes campos para cada película:

- título: Título de la película en castellano.
- título_original: Título original de la película
- año: Año de estreno de la película
- duración: Duración de la película en minutos
- país: País de producción de la película
- dirección: Directores de la película
- guion: Guionistas de la película
- reparto: Actores y actrices de la película
- productora: Productora de la película
- género: Géneros a los que pertenece la película
- sinopsis: Breve resumen de la película
- premios: Lista de premios que ha recibido
- puntuación_fa: Puntuación media recibida en Filmaffinity
- votos_fa: Número de votos recibidos en Filmaffinity
- puntuación_positiva: Número de críticas positivas recibidas de los profesionales
- puntuación_neutral: Número de críticas neutrales recibidas de los profesionales
- puntuación_negativa: Número de críticas negativas recibidas de los profesionales
- portada_url: URL del cartel de la película
- portada_local: Dirección local donde se ha descargado el cartel de la película

- puntuación_imbd: Valoración media recibida en IMDb
- votos_imdb: Número de votos recibidos en IMDb

El periodo de tiempo es toda la historia de Filmaffinity (desde 2002) e IMDb (desde 1990). La mayor parte de los campos que hemos recopilado no cambiarán con el tiempo, ya que contienen información de cada una de las películas. En cambio, los campos que informan sobre la puntuación y el número de votos, tanto en Filmaffinity como en IMDb, sí pueden cambiar con el tiempo, conforme los usuarios vayan registrando sus votos. Podemos suponer que aquellas películas con un número de votos muy grande serán menos propensas a cambiar su puntuación, a diferencia de las que tengan un número de votos pequeño, por ejemplo, las películas que se hayan estrenado recientemente.

Los datos se han recogido mediante web scraping en Python, accediendo a la página web de Filmaffinity para extraer la información general de las películas que se encuentran en el Top FA, su puntuación media y el número de votos de los usuarios. También se ha accedido a la página web de IMDB para buscar cada una de las películas anteriores y extraer de aquí la puntuación media y el número de votos de los usuarios de esta plataforma.

Para realizar todas estas tareas de web scraping se han usado varias librerías de Python:

- Selenium: Para recorrer la lista de las mejores películas de Filmaffinity y extraer la URL de cada una de ellas. También se ha utilizado para acceder al buscador de IMDb y buscar las películas de las que se quiere extraer la puntuación.
- Requests: Esta librería se ha utilizado para realizar una petición HTTP de la URL de cada película y guardar la respuesta.
- BeautifulSoap: Se ha utilizado para recorrer el DOM de la página web y extraer fácilmente la información deseada.

6. Agradecimientos

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Los propietarios de los datos son las empresas Filmaffinity S.L. e IMDb.com, Inc. Los derechos de propiedad intelectual de las críticas corresponden a los correspondientes críticos y/o medios de comunicación de los que han sido extraídos. El copyright del poster, carátula, fotogramas, fotografías e imágenes de cada DVD, VOD, Blu-ray, tráiler y banda sonora original (BSO) pertenecen a las correspondientes productoras y/o distribuidoras.

En los archivos robots.txt no se indica la prohibición de extraer información de las películas ni sus puntuaciones. En el caso de la página web de IMDb es un poco más restrictivo, ya que, por ejemplo, no permite descargar las imágenes de las películas (*Disallow: /title/tt*/mediaviewer/rm*/tr*). En nuestro caso, sólo necesitamos acceder a la puntuación en esta web, por tanto, no es una restricción que afecte a este proyecto.

Agradecemos a Filmaffinity S.L. y a IMDb.com, Inc. la recopilación de toda la información y hacerla disponible al público de forma libre.

7. Inspiración

Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El conjunto de datos generado puede ser interesante para realizar estudios de diferente tipo. Se pueden realizar análisis de estadística descriptiva de puntuación o número de premios de películas por país de producción, por directores, por actores y actrices, o por décadas, por poner algunos ejemplos.

Además, tenemos la puntuación de cada película en una plataforma española (Filmaffinity) y también la puntuación en una plataforma internacional (IMDb), por tanto, podemos realizar análisis de valoración internacional frente a valoración en España, teniendo también en cuenta el número de votos. Se puede analizar qué características de las películas hace que reciba mejor valoración en una plataforma o en otra.

También puede ser interesante en el ámbito de la minería de datos, para realizar modelos predictivos para determinar los premios que recibirá una película en función de las críticas recibidas por los profesionales y de las puntuaciones recibidas por el público. O incluso predecir la valoración que tendrá por parte del público en función de la valoración de la crítica profesional.

Otro posible uso sería utilizar el conjunto de datos como base de un sistema de recomendación a partir de los atributos de la película.

8. Licencia

Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- **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 licencia escogida para la publicación de este conjunto de datos ha sido *CC BY-SA 4.0 License*. Los motivos de la elección de esta licencia se deben a la idoneidad de las cláusulas que esta presenta con el trabajo realizado:

- Se permite copiar y redistribuir el material en cualquier medio o formato, incluso para uso comercial. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento al autor original.
- Las modificaciones realizadas sobre el trabajo original deberán distribuirse sobre la misma licencia, indicando el nombre del creador del conjunto de datos original y los cambios realizados. Así se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original, además de permitir que continúe distribuyéndose bajo los mismos términos que el autor original planteó.

Para más información se puede visitar el enlace de esta licencia de Creative Commons:
https://creativecommons.org/licenses/by-sa/4.0/

9. Código

Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código fuente en lenguaje Python está disponible en GitHub en el siguiente enlace:

https://github.com/rcorvial/web_scraping

10. Dataset

Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Enlace: https://doi.org/10.5281/zenodo.4670450
DOI: 10.5281/zenodo.4670450

Contribuciones	Firma
Investigación previa	P.L.L R.C.A
Redacción de las respuestas	P.L.L R.C.A
Desarrollo código	P.L.L R.C.A

11. Bibliografía

- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- https://www.selenium.dev/docs/site/es/
- https://www.browserstack.com/guide/selenium-scroll-tutorial
- https://es.wikipedia.org/wiki/FilmAffinity
- https://es.wikipedia.org/wiki/Internet_Movie_Database