Práctica 1: Web Scraping

Autor/es:
Alejandro Gallardo Alberola
Juan Rodríguez Vega

Fecha:
Noviembre 2020

Contenido

Enunciado: Descripción de la práctica a realizar

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

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

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).

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

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

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

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

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

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

Solucionario

1. Contexto

FilmAffinity es básicamente una web de información y recomendación de películas y series que combina reseñas de los usuarios y profesionales formando así una sólida comunidad de cinéfilos, además de alojar información de más de 100.000 títulos y más de 100.000.000 de críticas. A través de Wikipedia podemos encontrar la siguiente información:

"FilmAffinity fue creado en Madrid en mayo del 2002 por el crítico de cine Pablo Kurt Verdú Schumann y el programador Daniel Nicolás. Desde un principio la página constaba de un sistema recomendador de películas llamado "Almas gemelas", el cual mostraba las personas más afines en función de las puntuaciones que se dan a las películas. Tres años después se lanzó la sección de críticas, en donde los usuarios expresaban su opinión sobre una película.

La edición americana de PC Magazine (abril de 2004) incluyó a FilmAffinity entre las Cien mejores webs por descubrir del mundo (100 webs que no sabías que se podía vivir sin ellas). En el mismo número, la consideraba la mejor web en la categoría de entretenimiento.

Es un sistema recomendador de cine, con una base de datos donde se encuentra la ficha completa (técnica y artística) de gran cantidad de películas, documentales, cortometrajes, mediometrajes y series de televisión."

Dentro del tema elegido, sin duda FilmAffinity está entre las mejores opciones para recopilar información sobre cine. En concreto, este dataset recoge información sobre las películas nominadas y ganadoras en los premios Óscar en la modalidad de mejor película. En su web, FilmAffinity ofrece todos los premios otorgados por la academia desde su primera edición el 16 de mayo de 1929.

(fuente: Wikipedia – Para más información: https://es.wikipedia.org/wiki/FilmAffinity*)*

2. Título

Datos y características sobre las películas premiadas y nominadas al Oscar a la mejor película desde que se creó el premio.

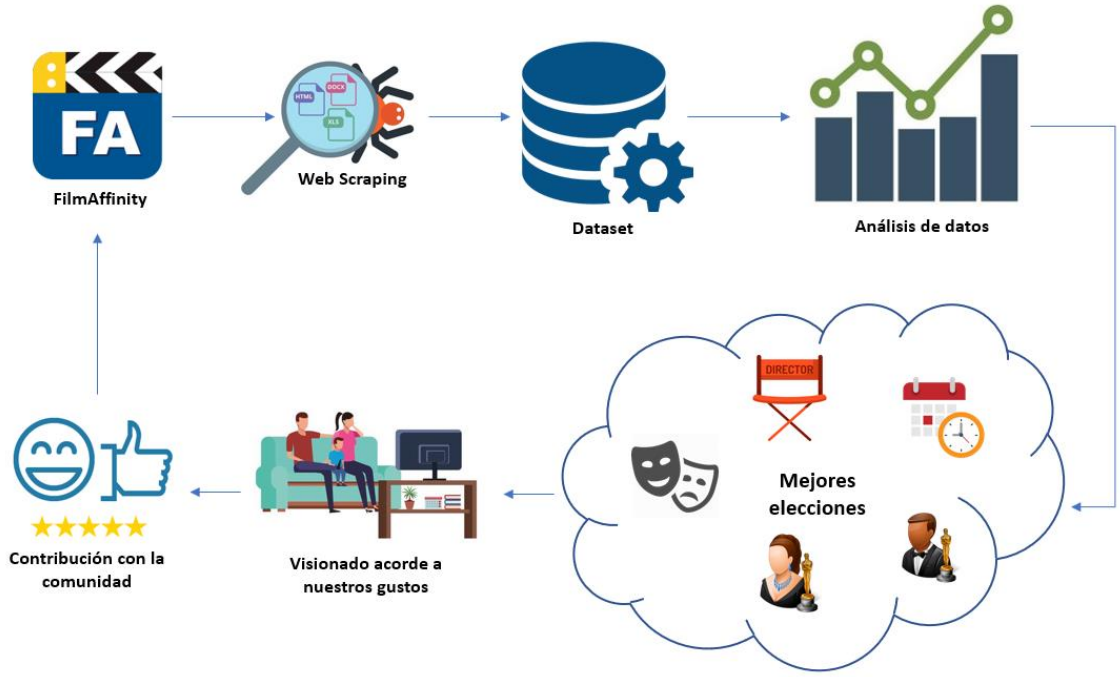
3. Descripción

Extracción de datos a través de técnicas de *web scraping* en la web *FilmAffinity,* relacionados con las películas nominadas y premiadas a mejor película en los Oscars y otorgados por la Academia de las Artes y las Ciencias Cinematográficas (AMPAS) desde la década de los años 20.

El dataset cuenta con información variada referente a las películas que alguna vez han optado al premio de la Academia como mejor película. Esta información se puede utilizar para clasificar estas películas entre las más vistas, las mejor valoradas, directores que han estado más veces nominados o que han ganado más premios. También actores que han participado en más ocasiones en películas premiadas y/o nominadas, géneros con mayor éxito en los Óscar, etc.

El data set se ha construido con todas las películas que han participado en el Oscar a mejor película desde que se creó el premio. Toda la información recogida se presenta en un fichero CSV que facilita su posterior tratamiento.

4. Representación gráfica



5. Contenido

Los campos de las películas premiadas y nominadas que se han extraído son:

- ***ID_FA:*** Identificador unívoco de FilmAffinity (extraído de la URL).
- ***title:*** Título de la película en castellano.
- ***original_Title:*** Título original de la película.
- ***year:*** Año de estreno de la película.
- ***duration:*** Duración de la película en minutos.
- ***country:*** País de producción de la película
- ***director:*** Director o directores cinematográficos de la película.
- ***actors:*** Listado de los actores principales de la película.
- ***genre:*** Lista de géneros en los que se puede enmarcar la película.
- ***description:*** Sinopsis de la película.
- ***awards:*** Premios y nominaciones de la película, no solo en los Óscar.
- ***average_rating:*** Puntuación media de las votaciones de la película en la plataforma.
- ***rating_votes:*** Número de votos válidos de la película por los usuarios de la web.
- ***reviews:*** número de críticas existentes sobre la película en la plataforma, tanto de usuarios profesionales como aficionados.
- ***poster:*** URL de la imagen de portada de la película

- ***my_vote:*** si la película ha sido visualizada y votada de forma personal por nosotros, registramos este voto dentro del dataset.

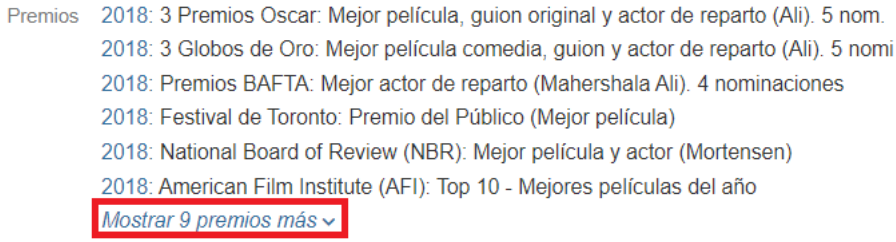
Todos estos campos son estáticos (siempre y cuando no haya erratas), a excepción de los campos *average_rating*, *rating_votes, reviews y my_vote*, los cuales resultan dinámicos y pueden ir variando con el tiempo en tanto los miembros de la comunidad hagan y registren sus valoraciones. También podría darse el caso menos probable de que alguna película fuese nominada o galardonada con un nuevo premio a posteriori.

Los datos han sido recogidos a través de web scraping en lenguaje Python. Se ha partido de una web en la que aparecen todas las películas nominadas y premiadas en la década de los años 2010 y se ha extraído la URL a cada una de las décadas en las que se han otorgado premios. Posteriormente, a raíz de la información por década, se ha obtenido la URL a la ficha de cada película individual, de la cual se ha extraído toda la información que se indicaba anteriormente. Por último, se ha descargado el poster o cartel de cada película y se ha pasado toda la información recopilada a un CSV.

Para la extracción se han utilizado ciertas tecnologías avanzadas que nos han facilitado ciertas tareas. Entre ellas, podemos destacar:

Tdqm: facilita ver de una forma visual el progreso de un bucle y, en nuestro caso concreto, del script de extracción de información.

Selenium: facilita la creación y edición de pruebas/acciones automatizadas de una manera muy sencilla. En nuestro caso particular, se ha utilizado para la extracción del listado de premios de las películas que vienen ocultos.



En términos prácticos, se debe comentar que no es estrictamente necesario utilizar Selenium para este caso concreto, puesto que los premios ocultos vienen incluidos (aunque especificados de una forma distinta a los que sí se muestran) en el código HTML que devuelve la petición a través de requests. No obstante, hemos decidido apoyarnos en dicha tecnología con fines didácticos, automatizando el *clic* en este elemento para el despliegue de los elementos ocultos y extraerlos directamente.

Este es un ejemplo de lo que permite hacer esta tecnología. Podríamos haberlo utilizado también, por ejemplo, para iniciar sesión con un usuario y contraseña de una forma dinámica y automatizada, aunque en nuestro caso no ha sido así porque hemos importado las cookies de sesión que ya se generaban en el *session requests.*

urllib.parse: módulo sencillo que facilita el manejo de URLs. En concreto, se utiliza en este proyecto para la gestión de URLs relativas y absolutas.

6. Agradecimientos

El propietario de los datos es la empresa FilmAffinity S.L. 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.

De todos ellos se hace un uso legítimo y razonable para fines académicos y/o informativos.

En el archivo robots.txt no se indica que se bloquee ninguna de las webs accedidas, por lo que el propietario del sitio no se opone de forma expresa a que se realice web scraping en su web.

La web ofrece algunas estadísticas referentes al número de premios por película o similares, pero realizando esta recolección de datos, sería posible realizar estadísticas y una extracción de datos mucho más profunda.

7. Inspiración

Este dataset puede resultar de interés para cualquier persona aficionada al cine, o para aquellas que lo están empezando a descubrir ahora. Dado que se registran películas ganadoras y nominadas al Óscar a la mejor película, es un contenido de calidad que puede ser consultado a la hora de querer buscar una buena película. Una cosa que ocurre con frecuencia es que la película ganadora del premio coge más fama que el resto, pero otras películas que solo han sido nominadas pasan más desapercibidas cuando también tienen una calidad cinematográfica excelente. El conjunto de datos también permite jugar con los diferentes atributos que tiene para sacar estadísticas muy interesantes en función de actores, director o géneros concretos, lo que nos permitiría buscar películas que se adapten fielmente a nuestros gustos.

Las preguntas que se pretenden responder son: ¿Qué películas de entre las nominadas o galardonadas tienen un mayor número de visualizaciones? ¿Y una mejor nota? ¿Puede influir el año en el que se estrenó en el número de visualizaciones? ¿Y el país en el que se produjo? ¿Qué otras películas nominadas u oscarizadas de mi actor favorito han estado nominadas? ¿Existe una relación entre el género de la película y sus opciones de ganar el premio? Estos son algunos de los ejemplos de las preguntas que este dataset puede responder, y que puede ser muy interesante para cualquier persona interesada en el cine.

8. Licencia

Este *dataset* se ofrece bajo licencia *Released Under CC BY-SA 4.0 License*. Esta licencia ha sido la elegida ya que permite:

- Copiar y redistribuir el material en cualquier medio o formato para cualquier propósito, incluso comercial, lo que puede ayudar a distribuir la información debido al interés de empresas del sector y también aporta reconocimiento del creador.
- Cualquier modificación o trabajo derivado basado en dicho material debe ser ofrecido bajo los términos de la misma licencia, indicando el nombre del creador original y las modificaciones realizadas sobre la obra original que permiten diferenciar las contribuciones, permitiendo así que la obra continúe bajo los mismos términos.

Para más información, se puede visitar el siguiente enlace de la licencia de Creative Commons:

https://creativecommons.org/licenses/by-nc-sa/4.0/

9. Código

Se adjunta el código fuente en lenguaje Python en el proyecto de GitHub disponible en:

https://github.com/alegalalb1/Web_scraping

10. Dataset

El dataset resultante se ha subido a Zenodo, se indican a continuación los datos de interés:

- Enlace: http://doi.org/10.5281/zenodo.4244742
- DOI: 10.5281/zenodo.4244742

Firmas

Contribuciones	Nombre	Firma
Investigación previa	Alejandro Gallardo Alberola	X
	Juan Rodríguez Vega	X
Redacción de las respuestas	Alejandro Gallardo Alberola	X
	Juan Rodríguez Vega	X
Desarrollo código	Alejandro Gallardo Alberola	X
	Juan Rodríguez Vega	X