

Práctica 1. Web Scraping

Adam Kepa. 13 abr. 19

Contexto

Para la realización de esta práctica, se ha dividido el trabajo en tres bloques o fases:

- **Búsqueda de información.** En esta fase se ha estudiado distintas páginas web, así como la información que aportan para la creación de un dataset que pudiese ser interesante.
- **Estudio de viabilidad.** Se han tenido en cuenta factores como la navegabilidad de la página web, la estructura de los datos, y la variabilidad del contenido.
- **Desarrollo de la solución.**

Finalmente se ha optado por el desarrollo de un dataset a partir de información de películas de la página web Filmaffinity.com. Se ha elegido esta página web por distintos motivos:

1. Es una página de referencia, y proporciona información muy completa sobre el tema que trata (información de películas).
2. La navegabilidad entre distintas páginas se realiza de forma muy homogénea mediante parámetros en la URL.
3. La estructura de los datos es bastante homogénea, lo que facilita su extracción, además de poder reutilizar el código para la creación de distintos datasets a partir de datos de la misma página web.
4. El fichero robots.txt no indica ninguna restricción para poder recorrer la estructura de la página web mediante un *web crawler*.
5. Los datos son estáticos, y no deberían variar mucho con el tiempo. Únicamente la información relativa a la valoración de las películas es dinámica.

Descripción

El título elegido para el el dataset ese *“film-award-dataset”*, puesto que recopila información histórica relativa a películas galardonadas por algún gran premio o festival. Cada entrada del dataset se corresponde al ganador de algún premio cinematográfico y agrega datos relativos al premio o festival, a la película, y a la valoración de la película por parte de los usuarios de la página web.

Contenido y representación gráfica

El siguiente modelo representa de forma gráfica el contenido de cada entrada del dataset.

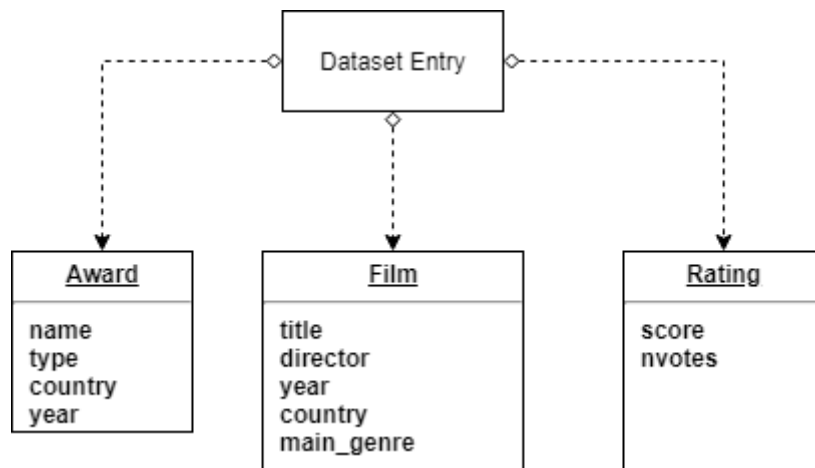


Figura 1. Estructura de una entrada del dataset

La descripción de estos campos es la siguiente:

- **award_name:**
 - Nombre del premio o festival.
- **award_type:**
 - Tipo de premio: “Premio”, “Festival”, “Premio asociación de críticos”
- **award_country:**
 - País en el que se celebra el premio o festival.
- **award_year:**
 - Año de edición del premio o festival.
- **film_title:**
 - Título de la película en español.
- **film_year:**
 - Año de estreno de la película
- **film_director:**
 - Nombre del director (o directores, separados por comas) de la película
- **film_country:**
 - País de origen de la película
- **film_main_genre:**
 - Género principal de la película
- **score:**
 - Nota media de la película, en base a los votos de los usuarios de Filmaffinity. Puede ser nulo en el caso de que la película no haya llegado a un número mínimo de votos.
- **nvotes:**
 - Número de votos en base a los cuales se ha calculado la nota. Puede ser nulo en el caso de que la película no haya llegado a un número mínimo de votos.

Los datos recopilados se corresponden a los disponibles en la página web el día 10 de abril de 2019. En el caso de que se ejecutase el script de forma posterior a esta fecha, se podría obtener un histórico actualizado de los premios, así como una versión actualizada de la valoración por parte de los usuarios.

Los datos han sido recopilados mediante técnicas de web scrapping, partiendo de la página principal de premios y festivales (https://www.filmaffinity.com/es/all_awards.php):

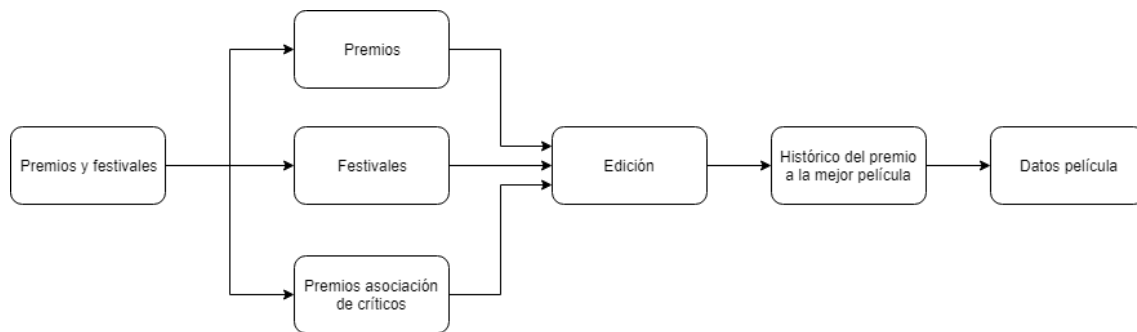


Figura 2. Navegabilidad para construir el set de datos.

Agradecimientos

Me gustaría expresar mi agradecimiento a los creadores Filmaffinity, así como a su comunidad de usuarios que hacen posible la existencia de esta página web.

Inspiración

El interés de este dataset radica en el hecho de que agrega datos de tres temas:

- Datos de un premio o festival
- Datos de una película
- Valoración por parte de los usuarios

Esto permite que el dataset pueda ser usado para responder preguntas de cualquiera de estos temas o las relaciones entre éstos. Permite encontrar la respuesta a preguntas como

- ¿Las películas de qué país han ganado en mayor número de ocasiones un festival?
- ¿Existe una diferencia significativa en la percepción por parte de los usuarios de las películas ganadoras de los festivales más importantes?
- ¿Existe correlación entre las películas ganadoras de un cierto premio, y el género de éstas?

Importando el dataset en Tableau podemos encontrar la respuesta a algunas de estas preguntas de forma muy sencilla:

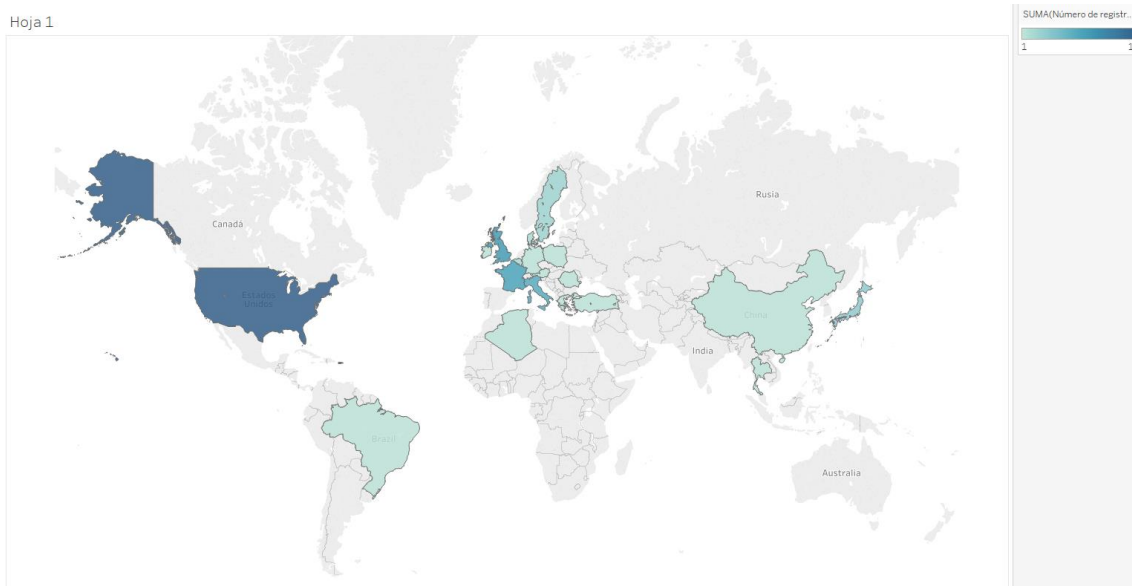


Figura 3. Países con mayor número de películas ganadoras del festival de Cannes. (Tableau)

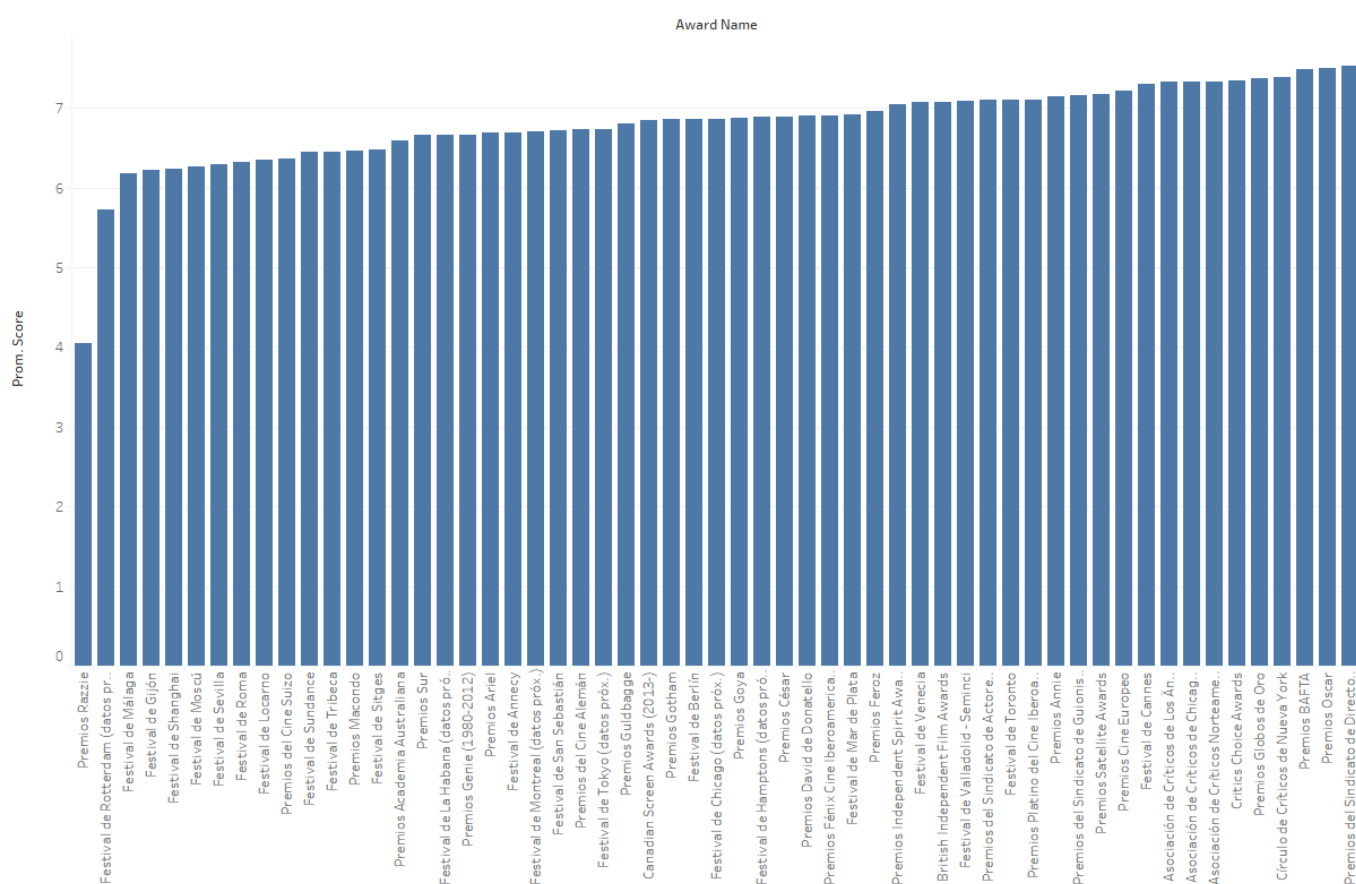


Figura 4. Valoración de los usuarios de las películas ganadoras de un premio o festival (Tableau)

Licencia

La licencia bajo la que se distribuiría el data set es **CC0 1.0 Universal (CC0 1.0)**.

Se ha optado por esta licencia debido al contexto educativo de esta práctica, así como el hecho de que el dueño de los datos extraídos es la página web Filmaffinity. Este tipo de licencia renuncia a tener ningún tipo de derecho de propiedad intelectual sobre los datos, y por tanto puede ser aprovechado en este mismo contexto para distribuirse y ser modificado sin restricciones, y sin la necesidad de mencionar al autor del dataset. Además, de esta forma el autor renuncia a ofrecer ninguna garantía sobre la obra.

Código

El código se encuentra en la carpeta /src del repositorio de Github.

Observaciones

La página web elegida tiene un mecanismo de protección contra bots que hace que, tras un número determinado de peticiones, las siguientes devuelvan un error HTTP 429 (Too Many Requests) hasta que se resuelva un Captcha. Se ha supuesto que la resolución de este problema queda fuera del alcance de esta práctica, y por lo cual el script se para hasta que este captcha se resuelve de forma manual.

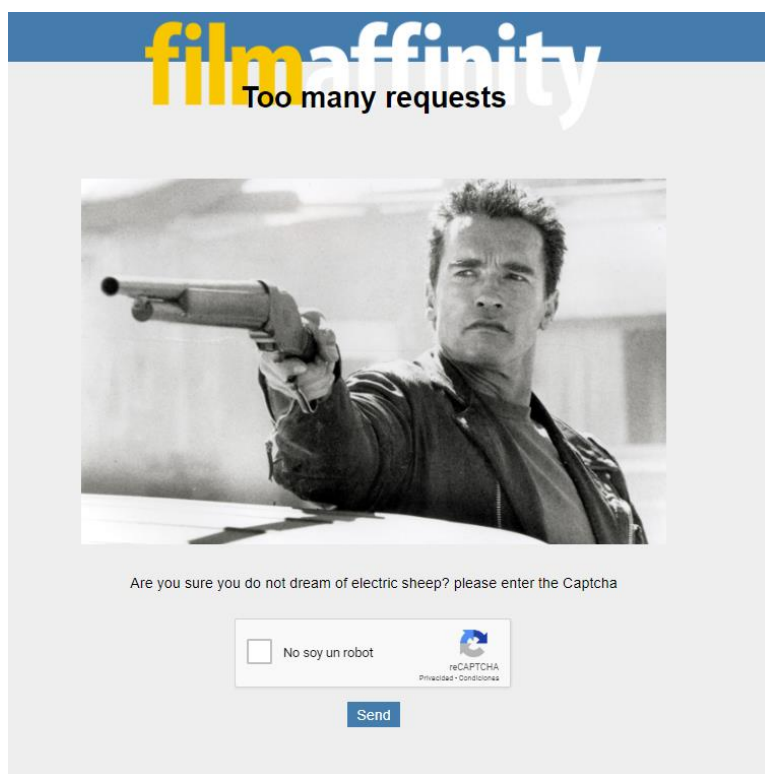
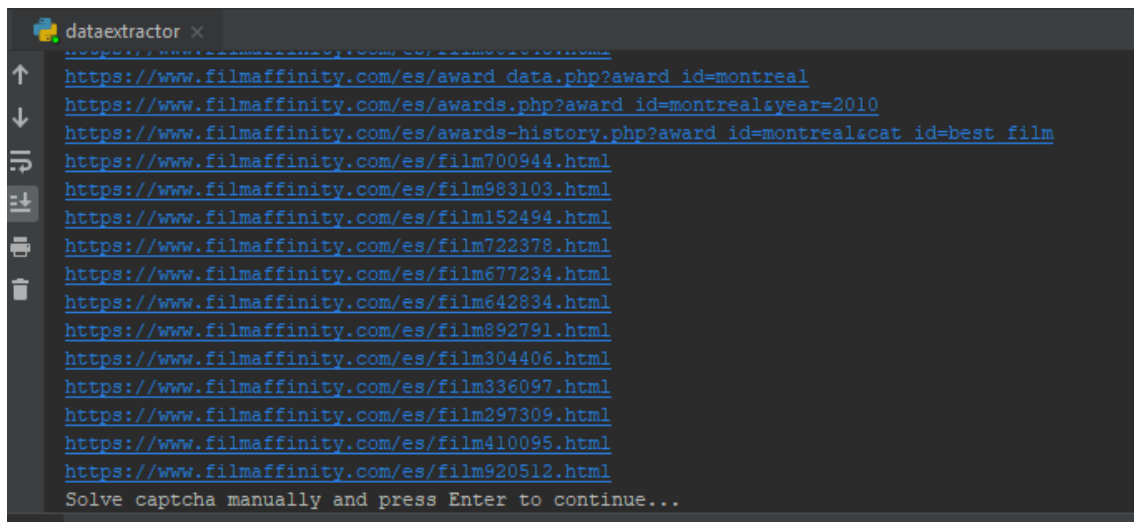


Figura 5. Captcha en la página de Filmaffinity



```
dataextractor x
https://www.filmaffinity.com/es/award_data.php?award_id=montreal
https://www.filmaffinity.com/es/awards.php?award_id=montreal&year=2010
https://www.filmaffinity.com/es/awards-history.php?award_id=montreal&cat_id=best_film
https://www.filmaffinity.com/es/film700944.html
https://www.filmaffinity.com/es/film983103.html
https://www.filmaffinity.com/es/film152494.html
https://www.filmaffinity.com/es/film722378.html
https://www.filmaffinity.com/es/film677234.html
https://www.filmaffinity.com/es/film642834.html
https://www.filmaffinity.com/es/film892791.html
https://www.filmaffinity.com/es/film304406.html
https://www.filmaffinity.com/es/film336097.html
https://www.filmaffinity.com/es/film297309.html
https://www.filmaffinity.com/es/film410095.html
https://www.filmaffinity.com/es/film920512.html
Solve captcha manually and press Enter to continue...
```

Figura 6. Parada del script, a la espera de resolver el captcha manualmente.

Dataset

El dataset se encuentra en la carpeta /csv del repositorio de Github.

Contribuciones	Firma
Investigación previa	Adam Kepa
Redacción de las respuestas	Adam Kepa
Desarrollo código	Adam Kepa