Creación de un dataset

Xavier Martínez Bartra

Martín Martínez Baltar

1. Contexto.

El **Internet Movie Database (IMDb)** contiene la base de datos de películas más grande y completa de la web. Esta web, pues, nos ofrece información extensa de todo tipo de películas de múltiples países y géneros, así como de series y contenido televisivo diverso. El sitio web fue fundado en 1990 y recientemente ha sido adquirido por Amazon. Cuando se quiere realizar indagación sobre cualquier ámbito filmográfico, la web ofrece información muy variada y densa de diversa índole, desde el género de las películas, su calificación, crítica, recaudación, director, actores, recaudación…. En el ámbito de extracción de información, pero, a parte de algunas fuentes de datos predefinidas, la web no ofrece un sistema de extracción automática de fuentes de datos en formato csv. o Excel de todos sus contenidos.

En este ámbito, en esta práctica, se ha realizado un scrapping del web de IMDB sobre las películas de ciencia ficción ordenadas por su recaudación. Para ajustar la recaudación de las películas a los datos del IPC, se ha ajustado la recaudación del listado de las películas al IPC con datos extraídos de la web del periódico económico Expansión. Este ajuste permite una comparativa sobre paridad de poder adquisitivo real de la recaudación de las películas.

(X.M.)
(M.M.B.)

2. Título.

Sci-Fi movies by inflation-adjusted gross revenue in the United States.

(X.M.)Ç
(M.M.B.)

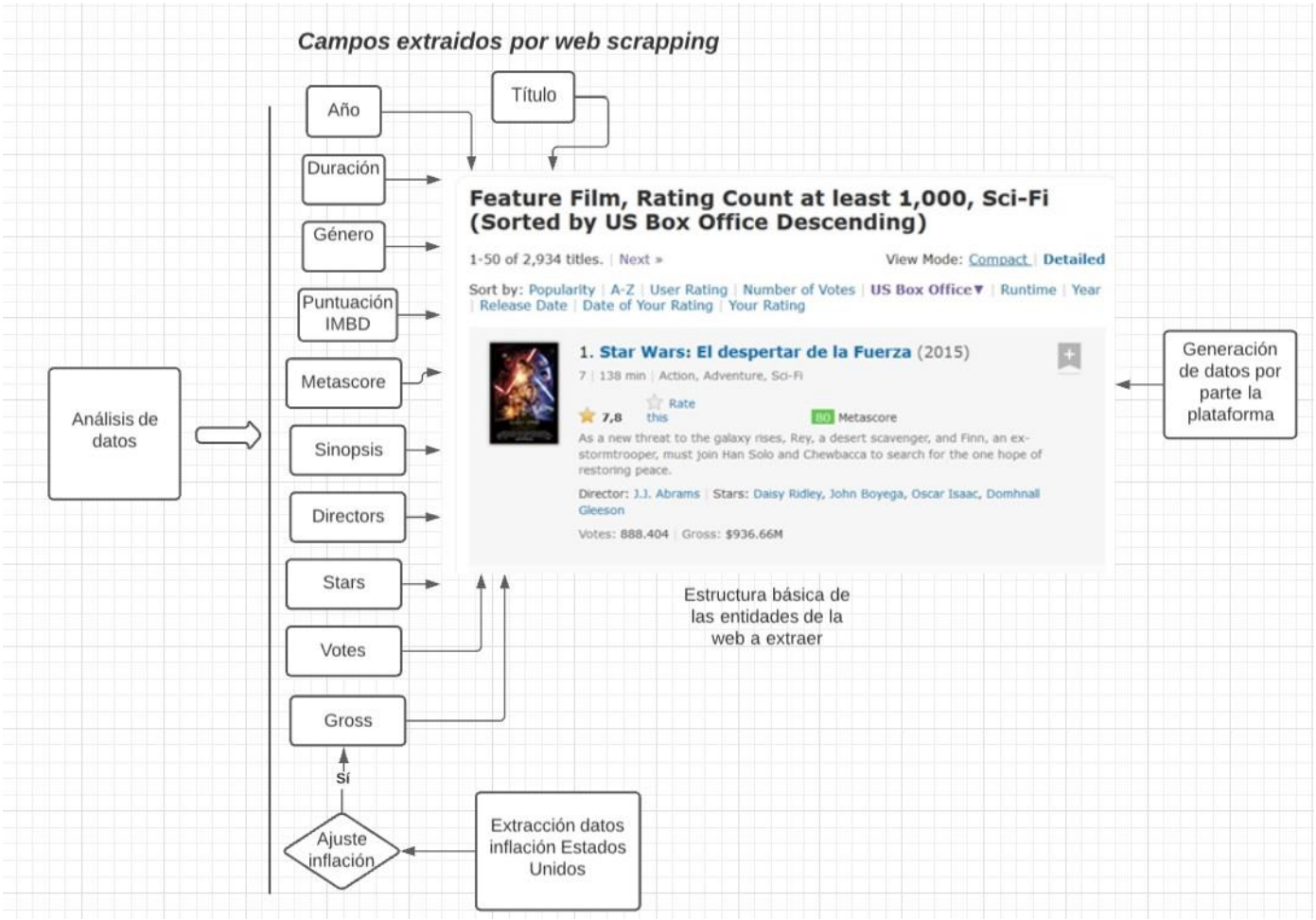
3. Descripción del dataset.

El conjunto de datos recoge los datos coleccionados por todas las películas de ciencia ficción del IMBD ordenadas por su recaudación en Estados Unidos. Los campos extraídos son **el título, el año, la duración de la película, el género, la puntuación dada por los usuarios de IMBD, la Metascore (nota de los críticos), la sinopsis de la película, el director, los actores, los votos totales que ha tenido en la web y finalmente la recaudación.** La recaudación se ajusta por la inflación utilizando los datos de la base de datos del periódico expansión.

(X.M)
(M.M.B.)

4. Representación gráfica.

El scrapping sigue el siguiente esquema básico para seleccionar las películas. Todas las películas siguen la misma estructuración básica :



(X.M.)

(M.M.B.)

5. Contenido.

El dataset incluye los datos de todas las películas de ciencia ficción con mas de 1.000 votos de los usuarios; ordenadas por su recaudación en la taquilla de Estados Unidos.

https://www.imdb.com/search/title/?title_type=feature&num_votes=1000,&genres=sci-fi&sort=boxoffice_gross_us,desc

El dataset incluye los siguientes campos:

- Title Título de la película
- Release_Year Año de lanzamiento
- Watchtime Duración
- Genre Todos los géneros incluidos
- Movie_Rating Ranking IMBD de la película
- Metascore Ranking de los críticos
- Votes Número de votos
- Gross_collection Recaudación
- Sinopsis Argumento
- Director Directores
- Star Actores
- Gross_equivalent Recaudación ajustando por inflación con los datos de https://datosmacro.expansion.com/

Los datos se han recopilado utilizando las siguientes librerías de scrapping de Python; Requests y Beautiful Soup. Hay que destacar que hay un número elevado de las películas (1821) que no tienen datos de recaudación. De todas películas solo se han extraído los datos presentes. Los años de las películas van del 1916 al 2021 y la película Star Wars: El despertar de la Fuerza es la de mayor recaudación ajustando por inflación. Ajustando por inflación, películas como ET extraterrestre de

1984 suben muchas posiciones en el ranking de recaudación de lo que sería con los datos nominales. *(X.M) (M.M.B.)*

6. Agradecimientos.

IMDB es una web que proporciona información sobre millones de películas y programas de televisión, así como todo tipo de información relacionada. El nombre es un acrónimo de Internet Movie Database. Como subsidiaria de propiedad absoluta de Amazon.com, IMDb tiene su sede en Seattle, pero la oficina de Col Needham, el fundador y director ejecutivo, permanece en Bristol, Inglaterra, donde se fundó el sitio web.

Realizar scrapping en esta web es muy popular entre los apasionados de la temática. IMBD establece que el público puede utilizar los datos de la web por uso no comerciales.

Para el scrapping de IMDb se pueden utilizar palabras clave para filtrar los datos de interés. IMDb siempre se actualiza con la información más reciente sobre cine y televisión. Algunos ejemplos de estudio cuantitativo de datos del IMBD es por ejemplo la tesis

KANIKA Almadi, Thesis: S.M. in Engineering and Management, Massachusetts Institute of Technology, System Design and Management Program, 2017 que utiliza los datos en bruto y no estructurados disponibles en el sitio web de IMDB, los limpia y los organiza en un formato estructurado adecuado para su análisis.

https://dspace.mit.edu/handle/1721.1/113502

En el ámbito más relacionado con el webscrapping, este enlace realiza todo tipo de técnicas de análisis exploratorio de datos.

https://towardsdatascience.com/data-analysis-and-visualization-of-scraped-data-from-imdb-with-r-5d75e8191fc0

(X.M)
(M.M.B.)

7. Inspiración.

El conjunto de datos obtenido nos permite realizar análisis exploratorio de datos sobre las películas de ciencia ficción extraídas, tanto desde el punto de vista de sus actores y directores, el tipo de sinopsis, la recaudación,….. La recaudación ajustada por inflación nos permite comparar de forma adecuada a través de los años películas como *E.T o Star Wars: El despertar de la Fuerza*; ya que nada tienen el valor de los dólares de la década de los 80 que en esta década.

Algunas de las cuestiones sobre las que podríamos indagar serían cuantas películas de ciencia ficción ha protagonizado cierto actor/director y que recaudación media han tenido sus películas.

También se podría realizar indagaciones NLP sobre sobre las sinopsis de las películas.

En el caso de la tesis del MIT, esta realiza minería de datos sobre los sets de sets de películas de IMDB y destaca varios parámetros como la interconexión de la película y las características del director, analizando si se correlacionan positivamente con la recaudación de la película. A partir de entonces, la tesis define vagamente un índice de innovación cinematográfica que abarca parámetros como el número de referencias, el número de seguidores y el número de remake y analiza cómo la abundancia de algunos de estos parámetros tiene un impacto positivo en el éxito de taquilla de la película.

(X.M.)
(M.M.B.)

8. Licencia.

Aplicar una licencia a los datos de su investigación ayuda a las personas que utilizan el juego de datos a comprender mejor como puede reproducirse la reutilización de los mismos y que tipo de cosas puede o no puede hacerse con ellos. Así, es muy recomendable que todo dataset contenga una licencia para los datos publicados. Los datos solo pueden usarse para uso personal y no comercial y no deben ser alterados / republicados / revendidos / reutilizados para crear ningún tipo de base de datos en línea / fuera de línea de información de películas excepto para uso personal individual.

La licencia elegida sería CC BY-NC-SA 4.0, que permite compartir, modificar los datos pero restringe su uso comercial y obliga el uso de la misma licencia.

La razón principal para elegir esta licencia es porque para construir el dataset hemos utilizado datos disponibles al público pero que son propiedad de terceros, por lo tanto, lo ético es permitir el acceso a estos datos como se nos permite a nosotros pero limitar el uso comercial que debe quedar reservado al propietario de los datos.

(X.M)
(M.M.B.)

9. Código:

El código se puede encontrar aquí.

10. Dataset:

El conjunto de datos generado se puede encontrar aquí.

11.Tabla contribución:

Contribución	Firmas
Investigación previa	Xavier Martínez Bartra, Martín Martínez Baltar
Redacción de las respuestas	Xavier Martínez Bartra, Martín Martínez Baltar
Desarrollo del código	Xavier Martínez Bartra, Martín Martínez Baltar