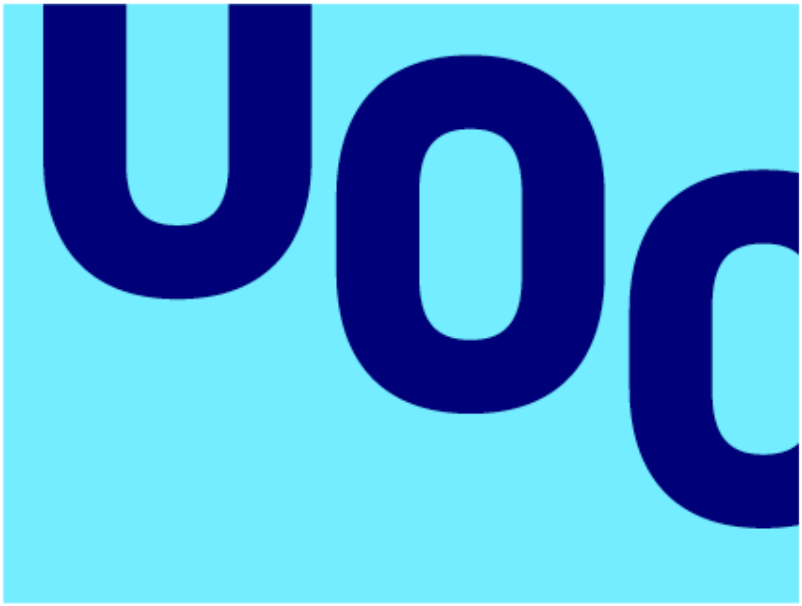

Universitat Oberta de Catalunya

M2.851 AULA 3

PRA 1

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

MASTER CIENCIA DE DATOS

Victor Mascarell Ascó

9 de noviembre de 2020

PRA1

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

Los datos del *dataset* se han recolectado a partir de la información, proporcionada por la **Wikipedia**, de las diferentes **películas estrenadas por Warner Bros durante los años 2010-2019**.

Se ha elegido Wikipedia como **fuente global y universal de información**, al tratarse de películas con copyright, bastante recientes, los departamentos de marketing de la empresa están pendientes que la información este correctamente actualizada en este sitio web.

Se ha elegido el período de **2010 a 2019, omitiendo explícitamente el año 2020**, por la situación global que vive la industria cinematográfica durante el mismo año. Debido al COVID-19, los datos de lanzamientos podrían verse sesgados respecto a otros años (reduciéndose drásticamente), lo que podría influir en las conclusiones de análisis posteriores.

Se ha elegido como fecha de inicio el 2010 porque en más de 9 años los hábitos de consumo pueden cambiar drásticamente (Un ejemplo es la inclusión en la industria cinematográfica de las nuevas plataformas de streaming como Netflix) y se pretende conocer cual es el comportamiento de los consumidores actuales.

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

"Peliculas Warner 2010-2019"

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 "*películas Warner 2010-2019*" ofrece **información sobre las películas lanzadas por la Warner Bross durante los años 2010 y 2019**. El cual cuenta **con 301 registros y 13 atributos**, entre los que se incluyen: fecha y día de lanzamiento, presupuesto y recaudación.

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

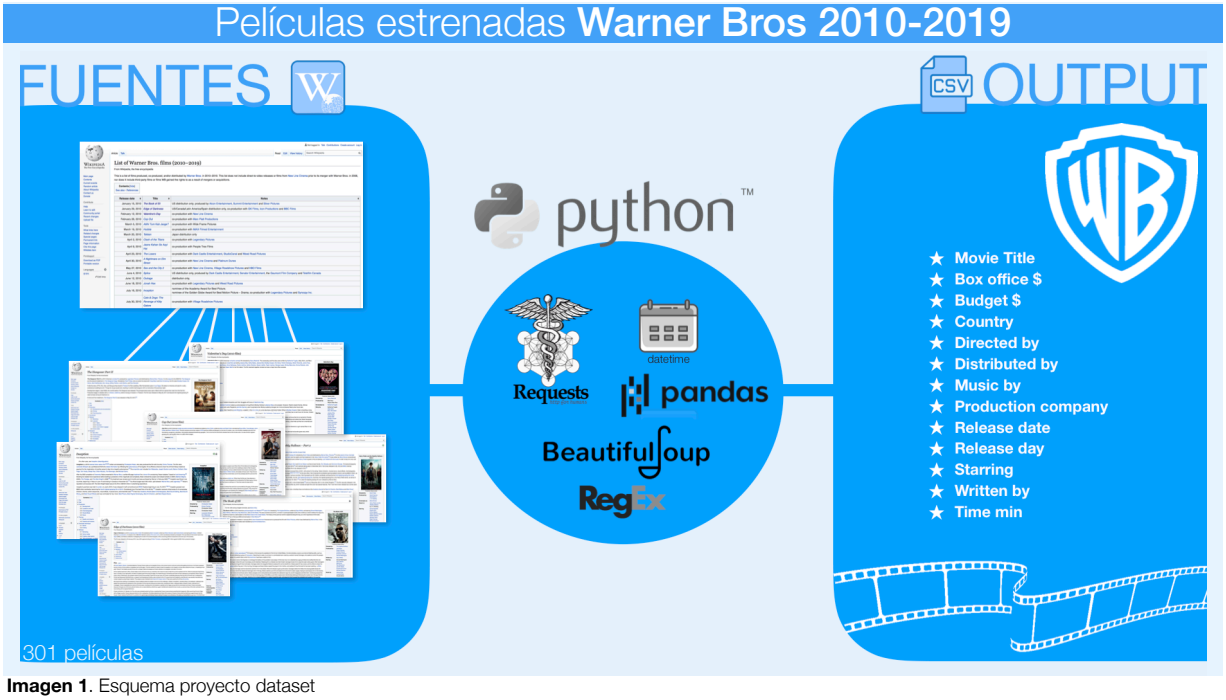

Imagen 1. Esquema proyecto dataset

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

El conjunto de datos cuenta con **301 registros** (películas) **y 13 campos**. Los campos incluidos son los siguientes:

- **Box office dolar:** Incluye la recaudación en dólares de la película. Tipo de dato: entero.
- **Budget dolar:** Incluye en presupuesto de la película en dólares. Tipo de dato: entero.
- **Country:** País responsable de la producción. Tipo de dato: string o lista.
- **Directed by:** Incluye a los responsables de la dirección de la película. Tipo de dato: string o lista.
- **Distributed by**: Incluye a los responsables de la distribución de la película. Tipo de dato: string o lista.
- **Music by**: Incluye a los responsables de la banda sonora de la película. Tipo de dato: string o lista.
- **Production company**: Incluye a los responsables de la producción de la película. Tipo de dato: string o lista.
- **Release dates:** Incluye la fecha de lanzamiento. En formato dd/mm/aaaa. Tipo de dato: string.
- **Release day:** Incluye el día de la semana en que se lanzo. En formato: Mon, Tue, Wed, Thu, Fri, Sat, y Sun. Tipo de dato: string.
- **Starring:** Incluye a los actores principales. Tipo de dato: string o lista.
- **Title:** Incluye el título de la película. Tipo de dato: string.
- **Written by:** Incluye a los responsables de escribir la película. Tipo de dato: string o lista.
- **Time min:** Incluye la duración de la película en minutos. Tipo de dato: entero.

El conjunto de datos cuenta con las **películas lanzadas por la Warner Bros** desde **2010 a 2019**.

¿Como se han recogido los datos?

Extracción

Inicialmente se ha tomado la web con el listado de películas estrenadas entre 2010-2019, para sacar los links de cada película (Ejemplo: "*The book of Eli*") y hacerles *web scraping* individualmente.

Las características de cada película se han obtenido de la caja que aparece a la derecha con la información principal. Para ello, se ha creado una **función que a partir de un link, obtuviera la información que queremos** y la incluyera en un diccionario.

Después se han **extraído los urls de cada película, del listado de películas**, y se han pasado en bucle para sacar la información y añadirla a una lista.

Limpieza básica

Se ha hecho una limpieza preliminar al sacar el contenido de la web, eliminando los superíndices y algunas etiquetas.Finalmente para un limpieza y unificación de los datos básica se han creado **4 funciones** que a partir de los datos obtenidos:

- **Unifican la moneda** (unificar_dinero_dolar (dinero)). A partir de un string, convierte todos datos a la misma moneda (Dólar) y en caso de estar en otro formato lo pasa a unidades en entero (ejemplos de otros formatos: €14 millon, $1.3 billion, $100,000, €300,000,…). Si hay una lista, se tiene en cuenta la primera entrada y si hay un rango ($14-15 million) se tiene en cuenta la cantidad inferior.
- **Unifican fecha de lanzamiento** (unificar_fecha_numerica(fecha)). Convierte todas las fechas al mismo formato: dd/mm/aaaa.
- **Sacan día de la semana** (sacar_dia(fecha)). A partir de la fecha de lanzamiento proporciona el día de la semana.
- **Transforman los minutos a entero** (minutos_a_int (duración)). A partir de un string con los minutos, elimina el texto y pasa los datos a entero.

Se han utilizado Python con las librerías: bs4 (BeautifulSoup), datetime, request, re y pandas.

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

El conjunto de datos "*peliculas Warner 2010-2019*" ha sido creado por **Victor Mascarell Ascó**. Se puede consultar el material consultado en el apartado de bibliografía.

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

Este conjunto de datos es interesante para hacer un **estudio sobre que día de la semana es el mejor para lanzar una película,** así como su fecha. Se podría analizar como influye el día de la semana en la recaudación de taquilla, así como si influyen otros aspectos a la hora de lanzar la misma. De la misma forma se puede estudiar que características tienen las películas que han apartado más beneficios (Recaudación - Presupuesto), para poder replicar el modelo en un futuro.

8. Licencia

El conjunto de datos "*peliculas Warner 2010-2019*" se encuentra bajo la **licencia CC BY-SA 4.0 License,** ya que mediante la misma, los datos pueden ser utilizados tanto a nivel comercial como particular, en cualquier medio o formato. Si se usan la mismos se debe referenciar al autor. Y en caso de utilizarlos, el material producido debe utilizar la misma licencia.
Además los datos de la Wikipedia también están bajo la misma licencia *Creative Commons Attribution-ShareAlike License.*

9. Código

El **código en python** (*peliculas_warner_10-19_scraper.py*) para generar el .csv puede **descargarse** en el siguiente enlace de GitHub: https://github.com/vicmascarell/peliculas-warner-2010-2019

10. Dataset

El ***dataset*** puede **descargarse** en el siguiente enlace de GitHub: https://github.com/vicmascarell/peliculas-warner-2010-2019

11. Contribuciones

CONTRIBUCIONES	FIRMA
Investigación previa	Victor Mascarell Ascó
Redacción de respuestas	Victor Mascarell Ascó
Desarrollo del código	Victor Mascarell Ascó

BIBLIOGRAFÍA

- Galli, K. (2020). *Comprehensive Python Beautiful Soup Web Scraping* [En línea]: https://www.youtube.com/watch?v=GjKQ6V_ViQE
- Galli, K. (2020). *Solving real world data science tasks with Python Beautiful Soup* [En línea]: https://www.youtube.com/watch?v=Ewgy-G9cmbg
- Masip, D. *El lenguaje Python*. Editorial UOC.
- Lawson, R. (2015). *Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data*.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons.
- Subirats, L., Calvo, M. (2018). *Web Scraping*. Editorial UOC.
- Tutorial de Github https://guides.github.com/activities/hello-world.
- Wikipedia (2020). L*ist of Warner Bros. films (2010–2019)* [En línea]: https://en.wikipedia.org/wiki/List_of_Warner_Bros._films_(2010–2019)