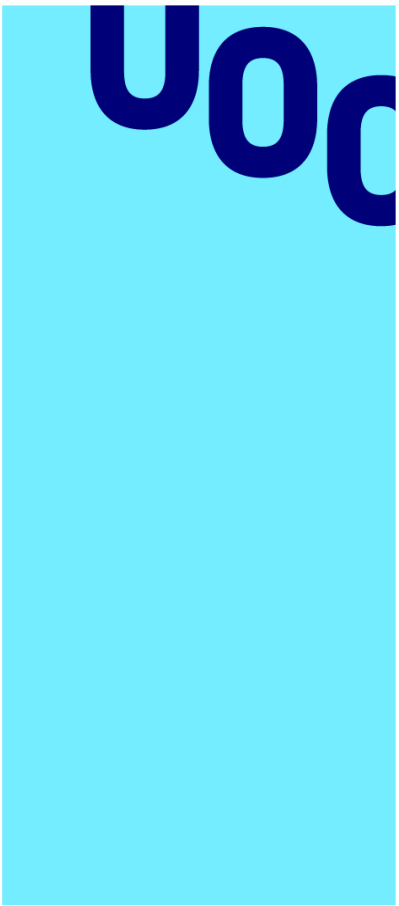
Tipología y ciclo de vida de los datos: Práctica 1

¿Cómo podemos capturar los datos de la web?

Agustin Rovira Quezada
Adrian Vega Morales



Universitat Oberta
de Catalunya

Índice

1. Contexto	2
2. Título	3
3. Descripción del dataset	3
4. Representación gráfica	4
5. Contenido	5
6. Propietario	6
7. Inspiracion	7
8. Licencia	7
9. Codigo	8
10. Dataset	10
11. Video	10
Contribucion	10
Bibliografía	11

1. Contexto

Como se ha observado a lo largo de los últimos años, la tendencia en consumo del contenido audiovisual de los españoles ha ido cambiando significativamente [1] y en especial tras la pandemia del virus conocido como COVID-19 [2]. Este último evento unido a la aparición de servicios streaming que ofrecen contenido audiovisual por catálogo bajo suscripción ha hecho que los espectadores migren de un consumo tradicional de televisión y cine a un consumo a demanda, en el que ellos eligen cuándo y qué ver, sin interrupciones.

Dentro de este marco, podemos observar que la extinción del cine tal y como lo conocemos está próxima. De esta forma, los cines han de adaptarse a las nuevas necesidades de los usuarios y al nuevo entorno tecnológico. No basta con proyectar las películas durante su estreno, sino que han de compatibilizar sus antiguos servicios con servicios a la carta para los consumidores de películas.

Esto podría ser tener un propio sistemas de streaming con el antiguo catálogo adquirido, hasta poseer un servicio de redes sociales en línea enfocado en compartir opiniones sobre las películas más interesantes.

Así, con el fin de acotar la idea y la realización de este trabajo, en este proyecto se propone hacer uso de las técnicas de web scraping para obtener una base de datos de películas que permita a posterior (out of scope) crear un sistema de recomendación de películas antiguas para cines, con el fin de re-monetizar y revalorizar clásicos del cine en la pantalla grande, Ej. Pulp-Fiction

Dado que la reproducción de películas en España pasa por entidades reguladoras que califican que se reproducen en los cines españoles, se consideró apropiado empezar tomando en cuenta para la creación del dataset *¿Que películas han sido aprobadas y reproducidas en los cines españoles?*. Esto por una parte, solo aportaría el catálogo de películas pero para que la base de datos sea relevante para el uso mencionado necesitamos completarla con información sobre actores, productor, recaudación, presupuesto, etc. Por esta razón elegimos 2 sitios webs de donde extraer la información

Sitio web 1: Catálogo de Cine Español del Ministerio de cultura y deporte español [3]
Catálogo de Cine Español integrado con la base de datos del ICAA de películas calificadas. Proporciona los títulos de las películas proyectadas en cines españoles

Sitio web 2: Catalogo de peliculas de Wikipedia
Contiene detalles como director, actores,etc de las películas. Además amplía el catálogo existente en la primera web, ya que no todas las películas fueron recogidas por el Ministerio de cultura y deporte español

2. Título

Filmografía occidental entre 2009-2015: Título y ficha técnica

3. Descripción del dataset

Como se menciona arriba, este dataset incluye los títulos de películas más su ficha técnica para el periodo comprendido entre el 2009 y el 2015. Se eligieron las películas pertenecientes a los mercados cinematográficos más influyentes del mundo occidental. Así, el dataset contiene películas de EE.UU, Francia, Italia, España, México, Argentina.

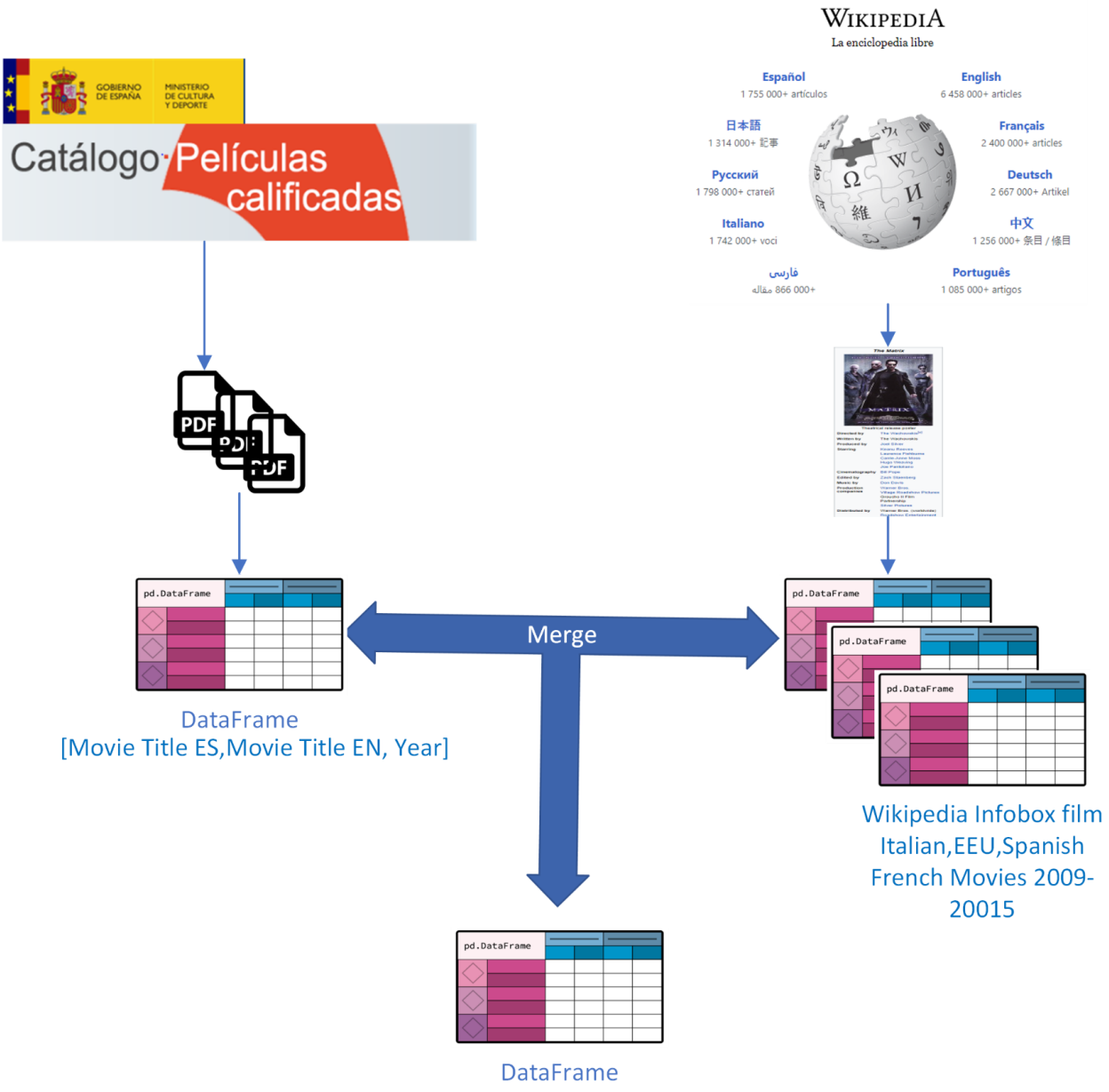
A cada fila del dataset le corresponde una película con tantos campos/columnas hay en la ficha técnica de la película. Por ejemplo, la fila 1 tendrá campos/columnas para *nombre de la película en español, nombre en inglés, año, dirigida por, actores, música,etc.*

Los campos del dataset en su mayoría son strings como por ejemplo el *nombre del director*, pero en otros casos son listas de strings como es el campo *Starring*. Hay unas pocas variables numéricas para el año de realización, de proyección, presupuesto y la recaudación de la película.

Por último cabe destacar que los datos no se han sometido a un preprocesamiento, por ello puede haber inconsistencias. Es necesario una fase de limpieza antes de poder hacer uso de los datos para un análisis.

El formato que se entrega este dataset es .csv para que su análisis y limpieza posterior se pueda hacer con librerías como Pandas de Python

4. Representación gráfica



[Movie Title ES,Movie Title EN, Year,Film name , director , writer
, screenplay , story , starring , language
, narrator,editing
, music , studio , distributor
, released , runtime , country
, budget , gross]

5. Contenido

Los datos corresponden a los títulos de películas más la ficha técnica de estos para el periodo comprendido entre el 2009 y 2015. Las filmografías seleccionadas se limitan a los países productores de películas más influyentes tales como EE.UU, Francia, Italia, España, México, Argentina.

Los datos fueron recolectados mediante Python, usando técnicas de web scraping. En una primera parte del código se implementa una navegación automática hasta llegar al catálogo de películas publicadas en el Ministerio de Cultura y Deporte. Luego se extrae automáticamente la información contenida en los PDFs que contienen el catálogo de películas para cada año.

Posteriormente, haciendo web scraping de Wikipedia, se genera automáticamente un dataset que contiene en cada fila un titulo de pelicula, el año, la URL y tantos campos como el infobox tenga wikipedia para esa película. Para ello primero recoge el título y la referencia en wikipedia de la película y posteriormente itera sobre cada una para completar la información con el infobox

Finalmente se hace un merge de los 2 dataset basándonos en el nombre de la película y el año. Se añade un campo más que hace referencia a si la película ha sido proyectada o no en un cine español.

Así el dataset final será guardado en un csv, *Filmografia occidental entre 2009-2015.* que debe contener siguientes campos:

- De la web del catálogo de películas publicadas en el Ministerio de Cultura y Deporte deben aparecer los campos:
    - spa_title: Título en español de la película
    - eng_title: Título en inglés de la película
    - year: Año de publicación de la película
- Del scraping de la sección de películas de Wikipedia se deberían sacar o debería contener, de acuerdo a la plantilla para *infobox* de películas de Wikipedia [8], los siguientes campos:
    - name: Nombre de la película.
    - image: Imagen significativa o poster de la película.
    - alt: Texto alternativo a la imagen.
    - caption: *Caption* de la imagen.
    - native_name: Título original de la película.
    - director: Director de la película.
    - writer: Guionista de la película.
    - screenplay: Director de escenografía.
    - story: Escritor de la historia.
    - based_on: Obra en la que está basada la película.

- Producer: Productor de la película.
- Starring: Elenco de actores/actrices.
- narrator: Narrador de la película.
- cinematography: Director de fotografía.
- editing: Editor de la película.
- music: Director musical.
- animator: Equipo de animadores o animador.
- layout_artist: Artista conceptual/Boceto
- background_artist: Artista de fondos y escenarios.
- color_process: Gama cromática de la película (Ej. blanco y negro, sepias, etc.)
- studio or production_companies: Estudio o productora.
- distributor: Distribuidora.
- released: Fecha de estreno.
- runtime: Duración de la película.
- country: País de producción.
- language: Lenguaje principal.
- budget: Presupuesto de la película.
- gross: Recaudación de la película.

Además de los campos que deben aparecer de acuerdo a esta plantilla encontramos ciertos campos como *spanish, catalan,* columnas con nombres casi idénticos o algún nombre que interpretamos como posibles erratas o inconsistencias en la estructura de las *infobox* en las páginas de las distintas películas y que por tanto tendremos que tratar en una fase de limpieza posterior.

6. Propietario

El dataset bebe de 2 fuentes como hemos comentado anteriormente. En el primer caso el propietario es el ministerio de cultura y deporte español, y en el segundo caso es Wikipedia.

En el primer no hemos encontrado análisis similares que respalden nuestra actuación, exceptuando por un post plantando How would you scrape this website using beautiful soup?. Aquí se describe cómo hacer web scraping al sitio web del Ministerio de Industria, Energía y Turismo.

Sin embargo para la página del ministerio de cultura y deporte, el propietario expresa que el contenido que se ofrece es meramente informativo y carece de efectos vinculantes para la Administración [3]. Además, al revisar el robots.txt de la página del ministerio [4] no encontramos ninguna restricción al uso de web scraping sobre dicha página web, por lo cual se creyó conveniente seguir el análisis llevado a cabo en este trabajo.

En el caso de wikipedia, la informacion extraida está sujeta a licencias CC BY-SA 3.0 lo cual hace necesario expresar explícitamente que parte del dataset ha sido obtenido desde Wikipedia. De hecho, más allá de eso no encontramos ninguna restricción al uso de técnicas de web scraping [5], es más, en el propio robots.txt menciona "*Friendly, low-speed bots are welcome viewing article pages, but not dynamically-generated pages please".*

Las inexistentes restricciones a la hora de realizar web scraping sobre wikipedia hacen que sea una página idónea para empezar a practicar dichas técnicas. Asi un análisis similar al realizado en esta práctica es el que encontramos en el siguiente post de medium:
Web Scraping Bollywood Filmographies on Wikipedia

7. Inspiracion

Con respecto al ejemplo anterior, hemos visto que se obtiene la información específica de películas de bollywood, y aunque este cine está en auge, muy pocas de estas películas han sido proyectadas en cine español o tiene doblaje al español. Por el contrario en nuestro análisis cotejamos la información de películas occidentales, que gozan de mayor fama, con la información aportada del ministerio de cultura y deporte, para así darle una información extra de que películas ya conocen los espectadores o ya han tenido la oportunidad de ver anunciadas.

El atractivo del análisis de este conjunto radica en que es posible encontrar patrones entre películas similares, bien ya sea basándose en el casting de actores, música, director. De esta forma estos patrones pueden ser usados en el cine para construir una red social de recomendaciones de películas y así incentivar a ver los nuevos estrenos ofertados o también para encontrar clásicos del cine que los espectadores estarían interesados en revivir en la pantalla grande.

De esta manera, el dataset podría dar respuesta a preguntas como, ¿Qué clásicos del cine puede revitalizar en pantalla grande? ¿Qué películas tuvieron un mejor ratio presupuesto/recaudación? ¿De los actores/director de las películas más rentables, que otras películas han protagonizado que sea de interés reavivar en el hype de dicho actor/director?

8. Licencia

Las fuentes de datos escogidas para construir el dataset, no presentan ninguna restricción a la hora de hacer el web scraping [4][5]. Sin embargo, al revisar el contenido de ambos sitios webs podemos ver que por un lado el ministerio de cultura y deporte ofrece el contenido de manera informativa y carece de efectos vinculantes para la Administración

pero por otro lado en wikipedia la mayoría de entradas se ofrecen bajo una licencia CC BY-SA 3.0 [6], permitiendo así a los usuarios el uso comercial de los contenidos reutilizados, siempre y cuando que los usos que se les den respeten las condiciones establecidas en la licencia respectiva.

Por lo tanto, la licencia que más se ajusta para la utilización de este dataset y el objetivo de este, es la CC BY-SA 4.0 License [7]. Es decir con esta licencia, siempre que otorgamos el crédito apropiado y sigamos los términos de la licencia, podemos:

- Compartir — copiar y redistribuir el material en cualquier medio o formato
- Adaptar: remezclar, transformar y construir sobre el material para cualquier fin, incluso comercial.

Estos dos puntos permiten que el dataset sea usado por otras empresas lo que daría pie a proyectos que reporten un reconocimiento al autor original.

9. Codigo

El código utilizado para la obtención del dataset es Python, haciendo uso de la librería beautifulsoup.

El código se encuentra ubicado en la carpeta source del repositorio webScrapingPeliculas. El link del repositorio: https://github.com/AdrianVega96/webScrapingPeliculas.git

Las librerías que se han usado para dicho proyecto están recogidas en un fichero .txt dentro del repositorio. Se podrán encontrar aquí.

Como veréis en la figura 2 el código tiene 2 líneas de procesos que dan lugar a 2 dataset que posteriormente cruzaremos para obtener el dataset final.

La primera línea se centra en la navegación del sitio web, usando beautifulsoup en un proceso iterativo llegamos hasta la página de descargas del catálogo. Posteriormente, *pdfURLs* se encarga de filtrar las URL de los archivos PDFs que contienen los catálogos de películas para cada año, las cuales se usarán de input en la siguiente función. En este punto llegamos a la función *getPDF* que se encarga de la extracción del PDF, página por página, línea a línea. Esta función tiene anidado un subproceso que limpia cada línea y filtra el nombre de las películas. Por último, el resultado es un dataframe con el título en español, inglés y el año.

La segunda línea de proceso se encarga de generar el dataset del catálogo de películas de wikipedia, más la ficha técnica de cada película. En primer lugar la función *Wikileaks* se

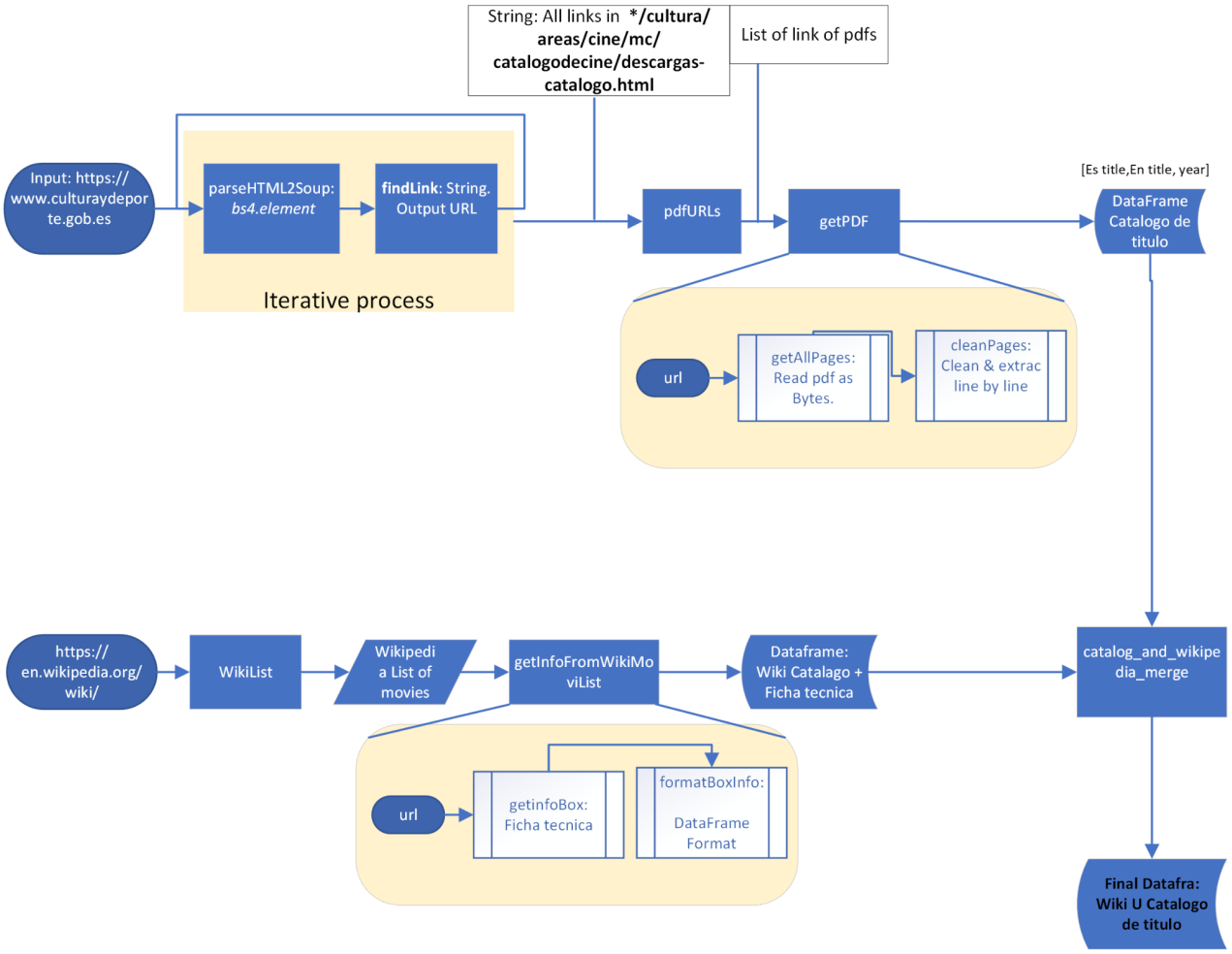
encarga de recolectar los títulos de películas más su referencia en wikipedia. Después, *getInfoFromWikiMoviList* se encarga de scrapear cada una de las referencias para encontrar la ficha técnica de la película. Así al final obtendremos un dataframe en el que cada fila es una película y las columnas son los campos técnicos como director, actores, etc.

Finalmente, los 2 dataset son unificados por una función *catalog_and_wikipedia_merge* que se encarga de hacer un merge de ambos basándose en el título y el año de la película.

El mayor reto que enfrentamos en esta práctica, aparte de familiarizarnos con los aspectos legales del web scraping y sus técnicas, ha sido la extracción y limpieza de de PDFs. Esta tarea ha requerido una investigación exhaustiva de los métodos actuales de extracción de PDFs ya que no hay un estándar o una solución general.

Además de esto, afrontamos una dificultad añadida en cuanto al formato del PDF. El Ministerio de Cultura y Deporte no sigue un patrón en el formato, lo cual hizo que se tuviese que combinar técnicas de regex con las librerías de edición de PDF de python para realizar un procesamiento personalizado.

A Continuación en la figura 2 podéis ver el flujo general del código.



10. Dataset

La URL de acceso al dataset subido en Zenodo es la que se muestra a continuación:

URL: https://doi.org/10.5281/zenodo.7337231

11. Video

El enlace al video es https://drive.google.com/file/d/1k_i36sAhEx6qQgaG4-qVpDCfZBv3xb3W/view?usp=sharing .

Además, subiremos el video a github por si surgieran alguna clase de problemas con los permisos de acceso.

Contribucion

La contribución a este proyecto se realizó a partes iguales entre los alumnos de la Universitat Oberta de Catalunya, Agustin Rovira y Adrian Vega.

Contribuciones	Firma
Investigación previa	Agustin Rovira, Adrian Vega
Desarrollo del código	Agustin Rovira, Adrian Vega
Redacción de las respuestas	Agustin Rovira, Adrian Vega
Participación en el vídeo	Agustin Rovira, Adrian Vega

Bibliografía

[1] POZO, C. (2022, February 10). *Cine en pandemia: recuperación, plataformas y alternativas*. RTVE.es. Retrieved November 16, 2022, from https://www.rtve.es/noticias/20220210/cine-pandemia-espana-recuperacion-plataformas-alternativas/2286905.shtml

[2] *Mapa del coronavirus en el mundo y datos de su evolución*. (2022, November 2). RTVE.es. Retrieved November 16, 2022, from https://www.rtve.es/noticias/20221102/mapa-mundial-del-coronavirus/1998143.shtml

[3] *Descargas del Catálogo de Cine Español - Catálogo - Películas calificadas*. (n.d.). Ministerio de Cultura y Deporte. Retrieved November 16, 2022, from

https://www.culturaydeporte.gob.es/cultura/areas/cine/mc/catalogodecine/descargas-catalogo.html

[4] Portada de la web del ministerio de cultura y deporte. (n.d.). Retrieved November 16, 2022, from https://www.culturaydeporte.gob.es/robots.txt

[5] *robots.txt on Wikipedia*. (n.d.). Wikipedia. Retrieved November 16, 2022, from https://en.wikipedia.org/robots.txt

[6] *Wikipedia:Derechos de autor*. (n.d.). Wikipedia. Retrieved November 16, 2022, from https://es.wikipedia.org/wiki/Wikipedia:Derechos_de_autor

[7] *Creative Commons — Attribution-ShareAlike 4.0 International — CC BY-SA 4.0*. (n.d.). Creative Commons. Retrieved November 16, 2022, from https://creativecommons.org/licenses/by-sa/4.0/

[8] Wood, E. D. (n.d.). Template:Infobox film. Wikipedia. Retrieved November 17, 2022, from https://en.wikipedia.org/wiki/Template:Infobox_film

.

.