PRÁCTICA 1 - web scraping

ANDER LOPETEGUI ARREGUI

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

Actualmente trabajo en una empresa que desarrolla webs para distribuidoras y productoras de cine donde se promocionan las películas y se realiza analítica de las visitas para aprovechar los datos de los usuarios en marketing. Las páginas webs ofrecen información sobre las películas, pero en muchos casos carecen de información sobre los actores. Se está pensando en añadir una sección con la lista de actores españoles con información básica y su foto. Por ello se ha decidido utilizar la información que proporciona la wikipedia. Como paso inicial se utilizará la página de categoría "Actrices españolas", donde se presenta una lista de links a cada página de la wiki de las actrices. Desde cada página obtendremos información básica de cada actriz. Con dicho dataset se podrá construir una página web con carteles con las fotos de las actrices y links a su propia wiki.

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

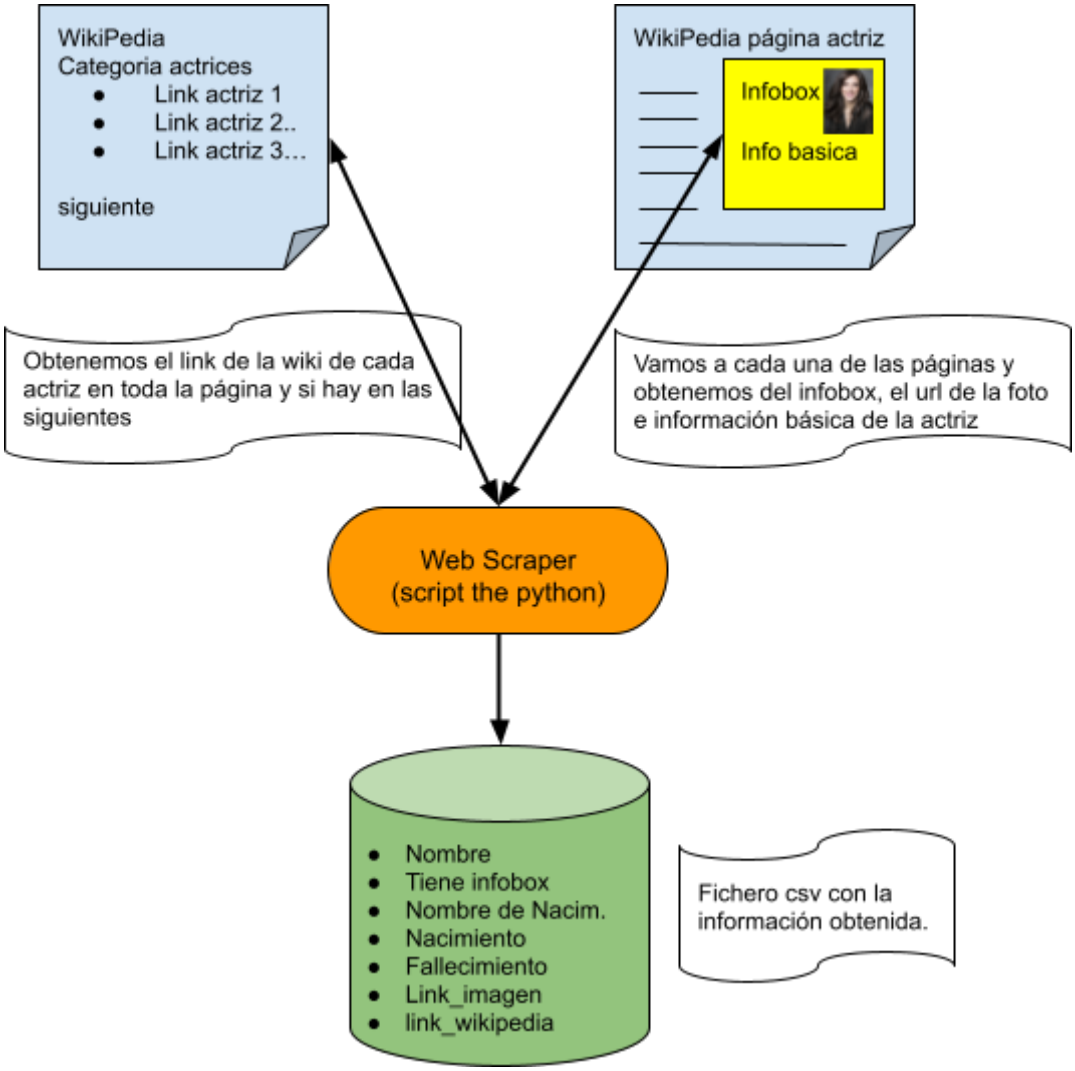
GALERÍA FOTOS ACTRICES ESPAÑOLAS

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 objetivo del dataset será construir una web con posters (fotos) de todas las actrices españolas con un subtítulo con información básica. Por ello el dataset contendrá el link de la foto de la actriz, su nombre artístico y su nombre de nacimiento. También información sobre el nacimiento y el fallecimiento si es que se ha producido.

Algunos campos relacionados con la información básica de la actriz pueden ser nulos debido a que su página de wikipedia no tiene infobox o le faltan algunos campos que queremos recoger. A pesar de que queremos realizar una galería de fotos, también guardaremos los datos de las actrices que no dispongan de foto para otros posibles usos.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



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

El dataset contiene los siguientes campos:

- **Nombre**: (texto). Nombre artístico de la actriz. Obtenido en la página de categorías.
- **Tiene infobox:** (booleano) sí o no tiene infobox. Obtenido en la página de la actriz
- **Nombre de nacimiento**: (texto). Obtenido dentro de la información de infobox de la página de la actriz.
- **Nacimiento**: (texto) Información sobre fecha y lugar de nacimiento. Obtenido dentro del infobox de la página de la actriz.
- **Fallecimiento**: (texto) Información sobre el fallecimiento de la actriz en el caso de que lo estuviese. Obtenido también del infobox.
- **Link_imagen**: (texto) dirección url de la imagen jpg de la actriz. Obtenido de la página de la actriz.
- **Link_wiki**: (texto) dirección url de la página de wikipedia de la actriz. Obtenido de la página de categoría de actrices.

El tiempo de los datos son duraderos porque no cambian a excepción de si ha fallecido y no tenemos la información actualizada. La wikipedia también puede quitar o modificar los enlaces url que tenemos, por lo que podrían dejar de funcionar.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

Para poder obtener toda esta cantidad de información y actualizada constantemente de una forma libre y gratuita no sería posible sin la wikipedia. No sería posible agradecer a algún contribuidor en particular, porque el scraping que estamos haciendo se realiza sobre un gran número de páginas diferentes y con autores diferentes. Por ello, mi principal agradecimiento a toda la comunidad de la wikipedia en general, que hace que esto sea posible.

Se pueden encontrar otros ejemplos de scraping de la wikipedia pero normalmente son realizadas a una única página de la wiki y no a una lista de páginas como es nuestro caso. Por ejemplo tenemos este repositorio de github donde se analizan los accidentes aereos en una página de wikipedia. https://github.com/anjanibhattar/Web-Scraping-Wiki-using-BS. O este otro ejemplo donde se el número de repeticiones de las palabras pero de nuevo en una única página https://github.com/junaidfiaz143/Wikipedia-Webscraping.

Al basarnos en datos que están publicados en la wikipedia, nos basamos en datos abiertos y así cumplimos con la legalidad y la LOPD. Los fines de este dataset simplemente son para unificar las fotos de las actrices y poder utilizarlas en una galería fotográfica, no modificando ni añadiendo datos que puedan incurrir en alguna ilegalidad. La ética del proyecto consiste en dar una información lo más actualizada posible y con la ayuda de la wikipedia podemos conseguirlo.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

La inspiración de crear este dataset viene dada por la necesitar de aglutinar toda la basta información que hay en internet en un único punto de una forma muy visual y centralizada. Existen comercializadoras que pueden ofrecerte esta misma información, quizás con fotos más profesionales de las actrices, pero siendo de pago y quizás no disponga de información de actrices del pasado y sólo tienen las actuales.

Considero que este dataset puede ser interesante porque en cualquier web de temática de cine pueden tener un apartado con esta galería, donde puedan permitir a sus visitantes repasar las fotos de las actrices del momento o del pasado, donde se pueda filtrar incluso for fecha de nacimiento para acotar búsquedas a las épocas que interesen.

Como he comentado, existen muchos web scrapers de páginas concretas de la wikipedia, pero no hay tantas o no he localizado aquellas que buscan en una categoría con listado de páginas. Además, el script creado permite introducir como parámetro la página de categorías, por lo que permite hacer scraping de cualquier categoría con estructura similar, como por ejemplo categoría de pintores, futbolistas, etc. Pudiendo crear otros datasets interesantes para otras temáticas.

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

La licencia seleccionada para el dataset es **CC BY-SA 3.0**. El motivo para elegir esta licencia se basa en las indicaciones de copyright que indica la propia wikipedia en este enlace: https://en.wikipedia.org/wiki/Wikipedia:FAQ/Copyright. En el cual se indican las licencias compatibles con Wikipedia y las que no son compatibles. Dentro del grupo de licencias *Creative Commons Licenses* la licencia CC BY-SA 3.0 es la que mejor se adapta a nuestro dataset. Debido a que el origen del dataset es de datos abiertos de wikipedia, consideramos que se debe continuar con una licencia “Free cultural”.

La licencia tiene las siguientes características:
Deja la libertad de:

- **Compartir** - Copiar y redistribuir el material en cualquier medio o formato
- **Adaptar** - Remezclar, transformar y crear a partir del material para cualquier finalidad, incluso comercial.

Bajo las siguientes condiciones:

- **Reconocimiento** - Se debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Se puede hacer de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.
- **Compartir por igual** - Si se remezcla, transforma o crea a partir del material, se deberá difundir sus contribuciones bajo la misma licencia que el original.

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

   Repositoría git público: https://github.com/alopetegui/wiki-scraper

10. Dataset. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

    Dataset Zenodo enlace DOI: https://doi.org/10.5281/zenodo.5644309

    Título del dataset: Spanish actresses photo gallery
    Fecha Publicación: 2021-11-04
    Licencia: Abierta, Creative Commons Attribution 3.0
    Tamaño: 75KB
    Número de filas: 405

Contribuciones:
Este proyecto ha sido realizado de forma individual.