PRÁCTICA 1
TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

MARÍA DEL PILAR FERNÁNDEZ FERNÁNDEZ
AULA 3

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

   El objetivo de esta recolección de datos es obtener información sobre los libros más vendidos en una de las principales webs de venta de libros (www.casadellibro.com). Esto nos permitirá obtener información de los libros que más demandan los lectores así como de la valoración que hacen de los mismos. Este conjunto de datos con otros obtenidos de webs similares nos permitirían obtener información sobre que libros son los mejor aceptados por los usuarios y esta información podría utilizarse entre otras posibilidades para webs de lectura donde se pueda hacer sugerencias de libros a lectores.
   La página web que queremos analizar es la siguiente: https://www.casadellibro.com/top-libros donde podemos encontrar los libros más vendidos de la casa del libro, separados por categorías. Para ello lo primero que haremos es analizar el archivo robots.txt donde observamos que no hay ninguna prohibición de acceso a este página por lo que podremos recolectar la información de la misma sin problemas.

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

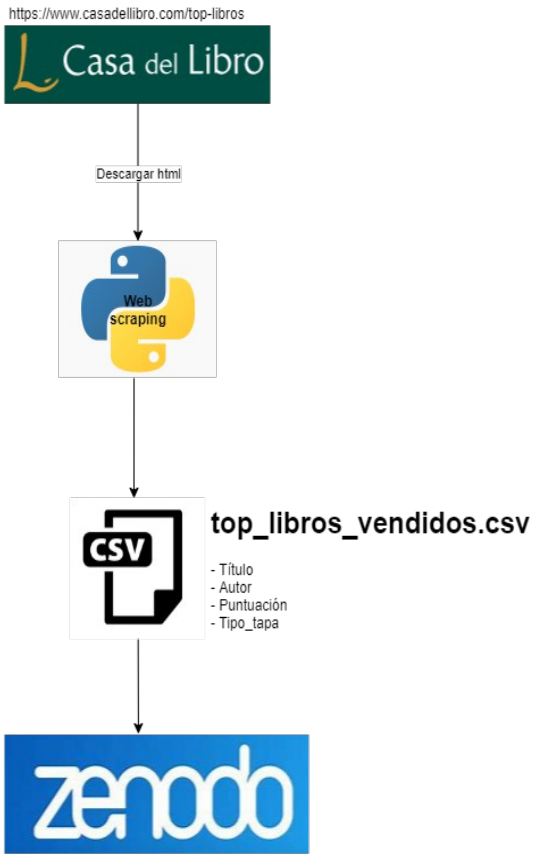
   El título que se ha elegido para el dataset es **top_libros_vendidos.csv** ya que contendrá información como se ha comentado anteriormente de los libros más vendidos.

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)**

   La recuperación de los datos referidos a los libros más vendidos en la casa del libro nos permitirá poder acceder a la información de los mismos y poder saber la puntuación que tienen los libros más leídos, cuáles son los autores de estos libros,...

4. **Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.**

   El diagrama muestra esquemáticamente los pasos realizados en el proyecto para extraer la información. En un primer lugar accedemos a la web de la casa del libro, a la sección de libros más vendidos, nos descargamos la página y realizamos web scraping con Python generando un dataset en csv con los datos que más nos interesan, una vez generado el dataset lo subimos a Zenodo para ponerlo a disposición de otros usuarios que quieran hacer uso de él.



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

   Los campos que contiene el dataset que se ha generado son los siguientes.
   - Título: se corresponde con el título del libro.
   - Autor: autor del libro.
   - Puntuación: puntuación asignada por los usuarios que han valorado el libro. La puntuación irá de 0 a 10 pudiendo encontrarnos con libros que todavía no han sido valorados en cuyo caso este campo vendrá vacío.
   - Tipo de tapa, pudiendo encontrar las siguientes opciones: tapa blanda, tapa dura y otras.

   La actualización de los datos de esta página según indica la propia página se hace mes a mes por lo que el dataset obtenido se corresponderá con la información que se encuentra en la web en el mes que se haya realizado el web scraping, si quisiéramos tener una imagen lo más cercana a la realidad debería realizarse una actualización mensualmente de los datos, o con más frecuencia si se detectara que la información cambiara más a menudo. La web de donde se ha recogido la información es la siguiente: https://www.casadellibro.com/top-libros

6. **Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.**

El propietario de los datos es casa del libro por lo que agradecemos que ofrezca la información para poder acceder a ella y tener una fotografía de cuáles son los libros más vendidos en todo momento. No he encontrado citas anteriores a estos datos. En cuanto a análisis similares podrían ser los que hacen páginas como letturalia que se encarga de buscar y comparar libros y precios entre diferentes webs pero no he encontrado información en la misma de los libros más vendidos de páginas web de venta de libros como la casa del libro, que es la que estamos tratando, o similares como la FNAC.

7. Inspiración. Explique 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.

Creo que este conjunto de datos es interesante para poder ofrecer información de los libros más vendidos y mejor puntuados para hacer sugerencias a lectores sobre ellos. Teniendo en cuenta que en este conjunto de libros también encontramos novedades se puede utilizar también para descubrir los libros más interesantes de entre los últimos que se han sacado al mercado.
Este conjunto de datos unido a conjuntos de datos de webs similares nos permitiría acotar aún más la información de los libros que ofrecer, ya que dispondríamos de la información de webs distintas que podrían estar orientadas a diferente tipo de público lo que nos permitiría poder ofrecer información sobre casos más específicos, por ejemplo que libro es el mejor valorado en varias webs. También podríamos buscar si algún autor se repite mucho entre los más vendidos, lo que podría ayudarnos a hacer sugerencias de libros relacionados con ese autor.

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**

1. Released Under CC0: Public Domain License.
2. Released Under CC BY-NC-SA 4.0 License
3. Released Under CC BY-SA 4.0 License
4. Database released under Open Database License, individual contents under Database Contents License
5. Other (specified above)
6. Unknown license

El dataset está publicado bajo la licencia CC BY-SA 4.0 License ya que esta licencia permite el uso por otros, incluso con propósitos comerciales siempre que se atribuya el crédito al creador del conjunto de datos generado.

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

El DOI asignado al dataset en Zenodo es el siguiente :10.5281/zenodo.4679655, allí se ha realizado una breve descripción del dataset subido

CONTRIBUCIONES	FIRMA
Investigación previa	María del Pilar Fernández Fernández
Redacción de las respuestas	María del Pilar Fernández Fernández
Desarrollo código	María del Pilar Fernández Fernández