Práctica 1. Web Scraping
Tipología y ciclo de vida de los datos
Universidad Abierta de Cataluña

Práctica 1. Web Scraping

Alumnas:
Raquel Martín de Consuegra Domínguez
Marta García González

1. **Título del dataset:** Ranking mejores películas
2. **Subtítulo del dataset:** Ranking con las películas mejor valoradas de todos los tiempos
3. **Imagen.**

4. **Contexto. ¿Cuál es la materia del conjunto de datos?**

Nuestro conjunto de datos está referido al ámbito del cine. Vamos a analizar conjuntos de datos con las películas mejor valoradas de todos los tiempos. Para ello hemos realizado dos arañas (imbd y ecartelera) para en un futuro contrastar los datos.

5. **Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?**

Como hemos creado dos arañas de distintas webs, cada una poseía datos distintos por lo que cada araña posee distintos campos.

- Araña imdb
  - puntuacionIMDB
  - generos
  - enlace
  - ranking
  - sinopsis
  - director
  - titulo
  - guionistas
  - anyoEstreno
  - duración

- Araña ecartelera:
  - ranking
  - genero
  - anyo
  - presupuesto
  - pais
  - tituloOriginal
  - duracion
  - titulo

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

Para realizar esta práctica hemos utilizado las web de imdb (http://www.imdb.com/) y ecartelera (https://www.ecartelera.com/).
Además nos hemos valido de distintos recursos para aprender a usar scrapy, como:

- Tutorial oficial de Scrapy: (https://doc.scrapy.org/en/latest/intro/tutorial.html#)
- Preguntas de stackoverflow: (https://stackoverflow.com)
- Licencias (http://www.gbif.es/gbif/ficheros/TallerDataPapers2013/C.b.Licencias-PM.pdf)
- Licencias (https://creativecommons.org)
- Youtube, con canales como Scrapinghub
- Uso de XPath (http://zvon.org/xxl/XPathTutorial/General/examples.html)
- Proyectos scrapy de github como (https://github.com/luisramirez-m/mercadolibre-scrapy)

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Este dataset puede ayudar a profesionales del sector cinematográfico a orientar su próxima producción. Pueden elegir los géneros, actores, etc. que más atraigan al público en función de un estudio de estos datos.

También puede generar interés en los cines ya que a través de los parámetros puede estimar la repercusión de las nuevas películas, de esta forma puede calcular el precio por el que puede ser rentable cada uno o qué prioridad en las salas pone a cada película.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

a. Released Under CC0: Public Domain License

Libera a nivel mundial y dentro de la ley, de todos sus derechos de propiedad intelectual, incluyendo todos los derechos conexos.

b. Released Under CC BY-NC-SA 4.0 License

Esta licencia obliga a reconocer el origen, a no comercializar con ella y a mantener esta licencia si se basan en ella. Se podrá compartir y adaptar.

c. Released Under CC BY-SA 4.0 License

Con esta licencia se podrá compartir y adaptar el trabajo, se ha de reconocer a la fuente y no se podrá comerciar con el mismo.

d. Database released under Open Database License, individual contents under Database Contents License

e. Other (specified above)

f. Unknown License

Elegimos el tipo de licencia b, Released Under CC BY-NC-SA 4.0 License puesto que los datos de IMDB y de ecartelera son de dominio público y disponibles para todos los usuarios.

Este tipo de licencia permite:

- Compartir, es decir, copiar y redistribuir el material en cualquier medio o formato.

- Adaptar, es decir, remezclar, transformar y crear a partir del material.
- El licenciador no puede revocar estas libertades mientras cumpla con los términos de la licencia.

Las condiciones que se deben respetar son las siguientes:

- Reconocimiento. Se debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios.
- El material no se podrá utilizar para una finalidad comercial.
- Si se mezclan los datos o se crean nuevos, se deberán distribuir mediante la misma licencia.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset.

Se puede ver en nuestro repositorio de github

10. Dataset: Dataset en formato CSV

Se puede ver en nuestro repositorio de github