Tipología y Ciclo de vida de los datos: Práctica 1

Master en Data Science

Tipología y Ciclo de vida
de los datos

Autor *Laura Teagno*

Fecha *Noviembre 2017*

Entrega *Práctica 1: Web Scraping*

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS – PRÁCTICA 1

Web Scraping

1. Título del dataset.
Películas más populares de España según IMDb.

2. Subtítulo del dataset.
Análisis de las características de las películas más populares de España.

3. Imagen.

4. Contexto.
El dataset se ha obtenido del análisis de la página web IMDb "Most Popular Titles With Country of Origin Spain". El objetivo es extraer las características de las películas más populares de España.

5. Contenido.
Las variables contenidas en el dataset son las siguientes:

- Posicion: Posición de la película del 1 al 50 de la lista de las 50 películas más populares de España hasta hoy.
- Titulo: Título de la película.
- Duración: Duración de la película.
- Genero: Genero de la película.
- Rating: Rating según IMDb de la película.
- Votos: Votos recibidos en favor de la película.
- Director: Principal director de la película. En caso de más de uno, se considerará el primero. En caso de ausente, se asignará el valor "NA".
- Actor: Principal actor de la película. En caso de más de uno, se considerará el primero.

Los datos han sido recogidos a través de web scraping con R. En particular con la función "rvest", comprobando el nombre CSS del elemento según el código html de la página web.

6. Agradecimientos.
Agradecer a Saurav Kaushik que en el artículo "Beginner's Guide on Web Scraping in R (using rvest) with hands-on example" explica muy bien cómo aplicar el scraping en las páginas web.

7. Inspiración.
Analizar cuáles son las películas más apreciadas por los españoles y sobre todo investigar si hay patrones repetidos, por ejemplo, si prefieren a los directores españoles o extranjeros, si de aventura o acción, etc.