Universitat Oberta de Catalunya

Máster Universitario en Ciencias de Datos

M2.851-Tipología y ciclo de vida de los datos - Aula 4

Práctica 1 [35 %]

Tema: Creación de un Dataset

Hubner Janampa Patilla - José Fernando Castillo A.
Estudiantes del Máster Universitario en Ciencias de Datos

1 de octubre del 2020

1. Competencias y objetivos

En esta Práctica se desarrollan las siguientes competencias del Master de Data Science:

1. Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
2. Capacidad para aplicar las técnicas específicas de web scraping.

2. Preguntas

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

El contexto para la recolección de información se realiza sobre el sitio web http://www.imdb.com/title/tt0944947/episodes, que nos muestra información sobre la serie **Game of Thrones**, IMDb es la fuente más popular y autorizada del mundo para contenido de películas, televisión y celebridades, diseñada para ayudar a los fanáticos a explorar el mundo de las películas y programas y decidir qué ver. Su base de datos tiene capacidad de búsqueda que incluye millones de películas, programas de televisión y entretenimiento. IMDb se lanzó en línea en 1990 y ha sido una subsidiaria de Amazon.com desde 1998.

El **Web Scraping** se realizó para Game of Thrones, que es una serie de televisión dramática de fantasía estadounidense creada por David Benioff y DB Weiss para HBO . Es una adaptación de Canción de hielo y fuego, la serie de novelas de fantasía de George RR Martin, la primera de las cuales es el Juego de tronos (1996). El programa fue producido y filmado en Belfast y en otras partes del Reino Unido. Los lugares de rodaje también incluyeron Canadá, Croacia, Islandia, Malta, Marruecos y España. La serie se estrenó en HBO en Estados Unidos el 17 de abril de 2011 y concluyó el 19 de mayo de 2019, con **73 episodios** transmitidos en **ocho temporadas**. Ambientada en los continentes ficticios de Westeros y Essos , Game of Thrones tiene un gran reparto y sigue varios arcos de la historia . Un arco trata sobre el Trono de Hierro de los Siete Reinos de Westeros y sigue una red de alianzas y conflictos entre las dinastías nobles, ya sea compitiendo por reclamar el trono o luchando por su independencia.

2.2. Definir un título para el dataset. Elegir un título que sea descriptivo.

El nombre definido para el DataSet es: **Game of Thrones**

2.3. 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).

Los campos extraídos mediante la técnica Web Scraping fue de la página web `http://www.imdb.com/title/tt0944947/episodes`, que corresponde a **Game of Thrones**, los campos extraídos fueron: *episodio*, *título*, *votos*, *rating* y *fecha de transmisión*, el escenario temporal de transmisión fue las ocho temporadas para los periodos de **abril del 2011 a mayo del 2019**.

2.4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

El cuadro 1 nos muestra como está estructurado el Dataset **GameofThrones.csv** para los 10 primeros registros:

Cuadro 1: Dataset Game of Thrones

Episodio	Titulo	Votos	Rating	F. Transmisión
1.1	Winter Is Coming	39834	9.1	17 Apr. 2011
1.2	The Kingsroad	30215	8.8	15 May 2011
1.3	Lord Snow	28567	8.7	1 May 2011
1.4	Cripples, Bastards, and Broken Things	27111	8.8	8 May 2011
1.5	The Wolf and the Lion	28220	9.1	15 May 2011
1.6	A Golden Crown	27925	9.2	22 May 2011
1.7	You Win or You Die	28419	9.2	29 May 2011
1.8	The Pointy End	26478	9.0	5 Jun. 2011
1.9	Baelor	37345	9.6	Jun. 2011
1.10	Fire and Blood	32781	9.5	19 Jun. 2011
...	...	...	...	...

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

Los campos extraídos de **Game of Thrones** fueron: *episodio*, *título*, *votos*, *rating* y *fecha de transmisión*, el escenario temporal de transmisión fue las ocho temporadas para los periodos de **abril del 2011** a **mayo del 2019**, a continuación, mostramos un extracto del dataset (Tabla 1) que consta de 67 filas y 5 columnas, que se obtuvo utilizando el Web Scraping sobre la página web descrita.

2.6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

En **IMDb** (Internet Movie Database), es una base de datos en línea que recopila información de películas, actores, series de televisión y otros temas relacionados con el mundo audiovisual. Amazon compra el 100 % de las acciones pasando a ser una empresa filial.

Extendemos nuestro agradecimiento a **IMDb** por disponer al alcance del público información útil, que en nuestro caso particular hemos aprovechado el sitio web para scrapear (Web

Scraping) y así obtener datos sobre una de las series más populares a nivel mundial. Por nuestra parte comunicamos que dicha información será utilizada con fines educativos y que estará a disposición de las personas que deseen utilizarla. El sitio web es un referente para obtener y analizar datos relacionados con el cine y los video juegos, por tal razón lo hemos tomado en cuenta para realizar la práctica.

A continuación, se detalla trabajos de investigación realizados en el sitio web:

1. Top 25 películas españolas. Enlace:https://www.imdb.com/list/ls039104853/
2. Investigar en el campo de los Estudios Fílmicos con bases de datos: AllMovie e IMDB. Enlace: https://www.lluiscodina.com/bases-datos-estudios-filmicos/
3. IMDb Top 250. Enlace: https://www.imdb.com/chart/top
4. Imdb.com Site Info. Enlace: https://www.alexa.com/siteinfo/imdb.com
5. El servicio de streaming IMDb TV: todo lo que te interesa conocer. Enlace: https://es.digitaltrends.com/entretenimiento/streaming-imdb-tv/

2.7. Inspiración. Explique por qué es interesante este conjunto de datos y que preguntas se pretenden responder.

El **Web Scraping** se realizó para **Game of Thrones**, que es una serie de televisión dramática de fantasía estadounidense, el dataset obtenido muestra información respecto a cada temporada y episodio, estos campos fueron: *episodio*, *título del episodio*, *los votos* de los fanáticos, el *rating* y la*fecha de transmisión*.

Las preguntas que pretendemos responder son:

1. ¿Qué temporada fue la más vista mediante la votación?
2. ¿Qué título de temporada y episodio tuvo el mayor y menor rating?
3. ¿Qué título de temporada y episodio tuvo la mayor y menor cantidad de votos?
4. ¿Fecha de transmisión por temporada y episodio?

2.8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección.

El dataset Game of Thrones Rating está bajo la licencia: https://creativecommons.org/licenses/by-nc-sa/4.0/. El mismo que estipula los siguientes términos de forma general en `Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)`, para mayor detalle revisar el link.

Hemos seleccionado la licencia `CC BY-NC-SA 4.0`, cuyo logo esta en la Figura 1, porque consideramos que todas las obras realizadas a partir del original ya sea para mejorar su composición, remezclar con otros datos o cualquiera que fuera su fin en los diferentes ámbitos como por ejemplo experimental o aprendizaje, se debe reconocer oportunamente el esfuerzo y dedicación que los autores principales han realizado para obtener dicho dataset. Además, desde el punto de vista comercial no se autoriza porque es un trabajo con fines educativos de cual sus autores han invertido tiempo y no se benefician de forma económica.

Por último, creemos que, si se llega a compartir o publicar la nueva obra con los nuevos cambios realizados, se debe distribuir su contribución bajo la misma licencia ya que se está permitiendo seguir con la idea de generar nuevo conocimiento a partir de las obras anteriores.

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

```
#Almacenar el sitio en una variable
url = 'http://www.imdb.com/title/tt0944947/episodes'

#Declarar los arreglos donde se guardarán los datos obtenidos
episodes = []
title = []
votes = []
ratings = []
dates = []

#Iteramos las temporadas de la serie (8 temporadas)
for season in range(1, 8):

#Obtener las peticiones del sitio web por temporadas
page = requests.get(url, params={'season': season})
soup = BeautifulSoup(page.text, 'html.parser')
pageEp = soup.find('div', class_='eplist')

#Iterar por cada temporada para obtener los datos de cada episodio
for epnro, div in enumerate(pageEp.find_all('div', recursive=False)):

#Se obtiene los episodios de cada temporada y se agrega al arreglo
episode = "{}.{}".format(season, epnro + 1)
episodes.append(episode)

#Se obtiene el título de los episodios y se agrega al arreglo
titles_t = div.find(itemprop='name')
titles = titles_t.get_text(strip=True)
title.append(titles)

#Se obtiene la votación de los episodios, se elimina los paréntesis,
# la coma y finalmente se agrega al arreglo
vote_v = div.find(class_='ipl-rating-star__total-votes')
vote = vote_v.get_text(strip=True)
vote = vote.lstrip('(')
vote = vote.rstrip(')')
vote = vote.replace(',', "")
votes.append(vote)

#Se obtiene el rating de los episodios y se agrega al arreglo
rating_r = div.find(class_='ipl-rating-star__rating')
rating = float(rating_r.get_text(strip=True))
ratings.append(rating)

#Se obtiene las fechas de estreno de los episodios y se agrega al arreglo
date_f = div.find(class_='airdate')
date = date_f.get_text(strip=True)
dates.append(date)

#Se guarda los datos en un Data Frame
dframe = pd.DataFrame({'Episodio':episodes,'Titulo':title,'Votos':votes,
'Rating':ratings,'F.Transmisión':dates})
```

```
print(dframe)

#Se exporta el Data Frame a un archivo excel y csv
dframe.to_excel('GameOfThrones.xlsx',index=False)
dframe.to_csv('Game_Of_Thrones.csv',index=False)
```

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

El dataset **Game of Thrones** está disponible en el **repositorio Zenodo**. Para acceder se proporciona la siguiente información:

1. URL:https://doi.org/10.5281/zenodo.4123037
2. DOI: 10.5281/zenodo.4123037

3. Contribuciones

Contribuciones	Firma
Investigación previa	Hubner Janampa -Jóse F. Castillo
Redacción de las respuestas	Hubner Janampa -Jóse F. Castillo
Desarrollo código	Hubner Janampa -Jóse F. Castillo

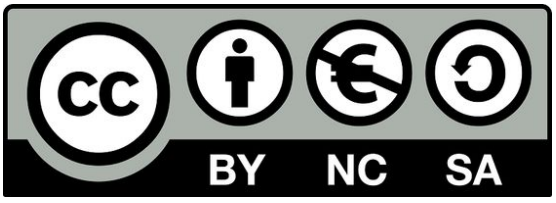

Figura 1: Licencia: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)