M2.851 Tipología y Ciclo de Vida de los Datos

PRA 1: Web Scraping

Daniel González Rodríguez

Contenido

1. Contexto

Dado el creciente desarrollo de contenido audiovisual en formato de series, y su popularización debido a las plataformas de *streaming,* se ha querido realizar este proyecto con el fin de recopilar un histórico de las mismas, y de esta manera poder analizar a futuro cómo ha evolucionado su producción.

Para ello se ha elegido la web TV Calendar (https://www.pogdesign.co.uk/cat/), ya que se trata de un calendario donde para cada día del mes se muestran las series de emisión, y por lo tanto permite recoger la información diaria de las series emitidas recorriendo desde el día actual hacia atrás, día tras día, hasta la primera emisión registrada en 1959. De esta manera se pueden agrupar los datos de cada serie, con sus correspondientes capítulos y fechas de emisión, además de recopilar las características de las mismas, para poder realizar análisis no solo de volúmenes de emisión, sino de sus características.

2. Título

El nombre elegido para el *dataset* es sencillamente "Series_DB", al tratarse de una base de datos de series, puesto que contiene un registro histórico de las series emitidas, manteniendo un registro de la fecha de emisión de cada uno de los capítulos de la misma.

3. Descripción del *dataset*

Ya que el histórico de una serie será la representación del total de todos los episodios emitidos en la misma, en el conjunto de datos que se ha extraído, se tiene que para cada fila existen los datos generales de la serie, y los específicos a cada capítulo distinto emitido en ella.

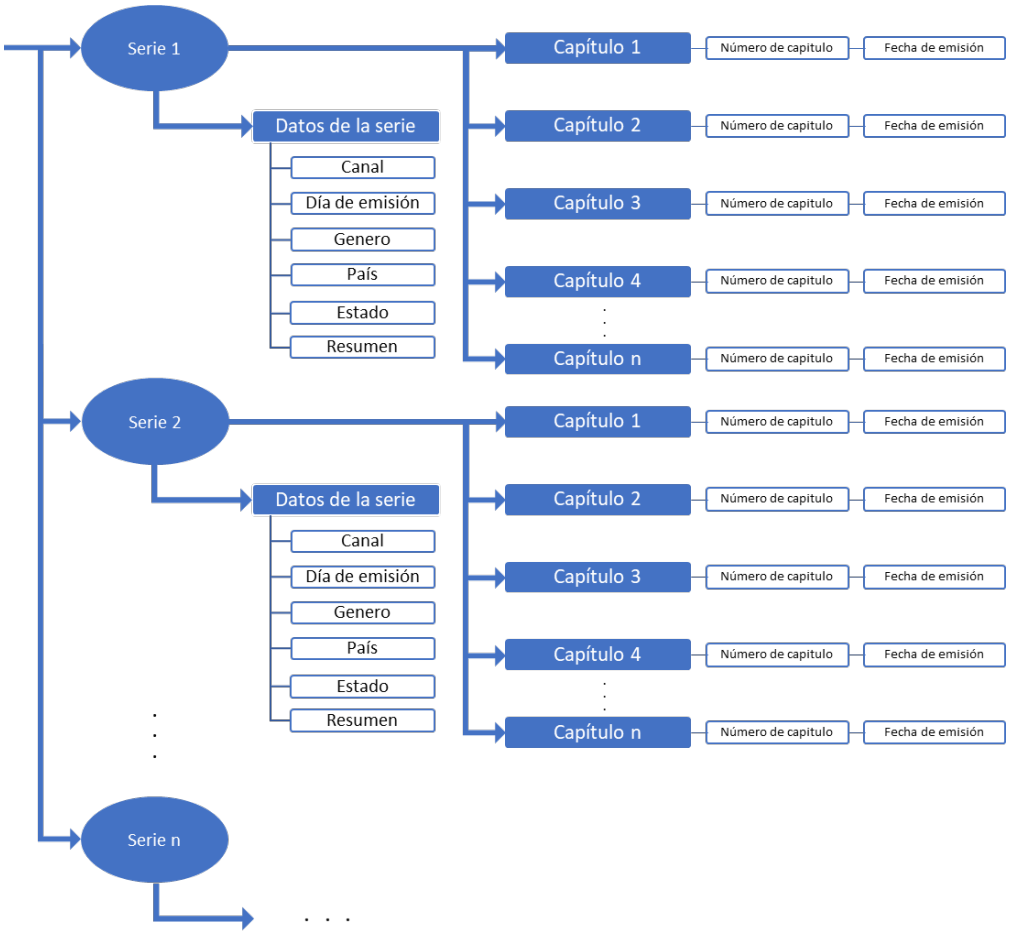
Así para cada serie se ha registrado cada capítulo perteneciente a la misma, con su correspondiente fecha de emisión y nombre del capítulo, lo que permite conocer todos los capítulos de cada serie, y además se complementa con información adicional de sus características, como género, día de emisión, etc., creando así una autentica base de datos de series.

4. Representación gráfica

En primer lugar, se muestra una captura de la web que ha servido como objeto del proyecto, para que se pueda dar lugar a visualizar como se encontraban los datos en bruto en la misma. Se puede observar como para el mes en cuestión, cada día contiene las series que se emitieron, y estas contienen enlaces a otras páginas que contienen la información de la serie. Además, se puede navegar hacía meses anteriores para poder automatizar el proceso de recolección de la información:

Fuente: *TV Calendar*[1]

Se representa visualmente como se construyen los datos que contiene el *dataset* en un formato grafico a modo de resumen. De esta manera se ejemplifica más fácilmente como se estructuran los datos de manera que se puedan entender cómo se distribuyen en el *dataset* donde se encuentran con una entrada para cada capítulo:



Fuente: *Elaboración propia*

[1] https://www.pogdesign.co.uk/cat/10-2000

5. Contenido

En el *dataset* se genera una entrada por cada capítulo emitido de la serie en cuestión, por lo que cada entrada contendrá los siguientes datos:

- **Serie**: Nombre de la serie.
- **Network (Channel)**: Canal donde se emite o emitió.
- **Category / Genre**: Genero al que pertenece (comedia, drama, etc.).
- **Broadcast Airs**: Franja horaria en la que se emite.
- **Country**: País de origen.
- **Episode Length**: Duración de los episodios.
- **Show Status**: Estado actual de la serie (finalizada, renovada por otra temporada, etc.)
- **Summary**: Resumen de la serie.
- **Episode Number**: Número de episodio en formato número de la temporada y capitulo de la temporada.
- **Aired Date**: Fecha de emisión del capítulo.

Así el *dataset* contendrá datos desde la primera serie registrada en 1959 hasta la fecha presente de la última extracción. Al encontrarse los datos originales en un formato de calendario, donde para cada día se indican las series que se van a emitir (o emitieron), cada dato (en este caso cada capítulo) vendrá reflejado por una fecha en formato año-mes-día, siendo el día la mínima granularidad en la serie temporal.

De esta manera, en cada entrada diaria para cada serie se puede ver el capítulo de emisión, pero además contiene un enlace hacía los datos generales de las series, por lo que si para una serie en cuestión no se conocían los datos de la serie se navega a esa web, y se recupera la información general de la serie. Con estos datos, y una vez se hayan navegado por todos los días que se emitió la serie, se tendrán los datos completos de dicha serie.

6. Agradecimientos

Los datos originales pertenecen al *website* https://www.pogdesign.co.uk, pero no se ha encontrado ninguna referencia a su propiedad o licenciamiento, ni que pertenezca la web a una entidad empresarial o con ánimo de lucro.

Se localizó en *GitHub* un proyecto que también toma este *website* como base (https://github.com/rafakob/service.pogdesign.sync), pero cuya finalidad es tratar una base de datos propia de series con el fin de actualizar las series que se han visto. Aunque los accesos se realizarán a los distintos capítulos no trata de obtener una base datos histórica de los capítulos como en el *dataset* generado en este proyecto.

Existen otros *websites* que contienen bases de datos de series, como TV Series Database (https://www.imdb.com/list/ls067528614/), pero su objetivo suele ser el acceso a la información en si misma de cada serie, y no a la investigación y evolución del contenido audiovisual de este tipo, por lo que no presenta acceso al contenido en bruto, y también prohíben realizar *web scraping* sobre ellas, por lo que no se podría obtener el *dataset* que se ha generado en este proyecto.

En línea con esto último, para asegurar la legalidad de la tarea, se realizó una pequeña investigación para asegurar que todo se mantenía dentro de los estándares, siguiente los siguientes pasos:

- Se buscaron las condiciones de uso, pero no existen, y no se pueden observar así las prohibiciones específicas que realizan otras tantas webs. De esta forma solo se ha accedido a información pública que se podría visualizar navegando por la web como cualquier persona podría hacer.
- Se comprobó que no existía información respecto del archivo *robots.txt* que no permitiría navegar o recopilar datos por determinadas carpetas de la web.
- Existe numerosa información adicional con el detalle de cada capítulo, pero con el fin de mantener la política de "no hacer daño" se decidió no recuperar esta información, ya que por ejemplo el resumen de cada capítulo apenas daría información al futuro análisis y resultaría en una cantidad conexiones adicionales enorme hacia la web.
- En esta línea, se ha procedido a programar un temporizador para espaciar las peticiones temporalmente al servidor, de manera que se evite cualquier tipo de sobrecarga.

Así, se mantiene el proyecto dentro de las buenas prácticas del *web scraping*, de manera que el proceso de recopilación de la información no suponga un perjuicio para ninguna de las partes.

7. Inspiración

Dada la creciente tendencia que las personas tienen a emplear sus horas de ocio en visualizar series, resulta interesante analizar cómo ha crecido esta industria. Poder conocer simplemente como ha evolucionado el volumen de series creadas, dará datos de como la industria ha evolucionado con el paso de los años. Además, se pueden obtener características interesantes desde el punto de vista de los distintos géneros, y como han cambiado los gustos de la población desde preferir el genero *sit-com*, a series basadas en entornos más fantásticos.

Este *dataset* permite analizar datos en el vacío que dejan aquellos *websites* comentados en el apartado 6, donde tan solo muestran una información desde el punto del contenido de la serie en si misma, simplemente para el usuario que desea visualizarla, pero no permiten analizarlas desde un prisma superior, que es lo que pretende proporcionar este proyecto.

Así mismo, tener en un solo *dataset* toda la información de las series no solo permitiría realizar análisis de la evolución de la industria, sino también la construcción de sistemas de recomendación para el usuario. Ya que no solo se cuenta con el genero de las series, sino también con un resumen de la misma, por lo que empleando técnicas de NLP, se podrían obtener síntesis de que caracteriza el argumento de la serie, y así recomendar al usuario series parecidas según sus gustos.

8. Licencia

En este caso se ha seleccionado la licencia **CC BY-SA 4.0 License** para la publicación de los datos.

Esto se elegido así, ya que de esta manera se haga referencia al origen, y pueda siempre dar opción a contrastarse como se ha construido el *dataset*, y como se podría seguir actualizando a través del código puesto a disposición en el repositorio.

Este último punto enlaza con el otro fundamento de haber seleccionado este tipo de licencia, y es que así se puedan modificar los datos, manteniendo una referencia al trabajo original.

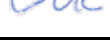
9. Código

Se adjunta el enlace al repositorio donde se encuentra todo el código generado para la recuperación de datos y construcción del *dataset*:

https://github.com/Dani643/WebScraping

10. Dataset

Se ha procedido a publicar el *dataset* en *Zenodo*, generando el siguiente DOI: **10.5281/zenodo.6426938**, cuya URL directa es: https://doi.org/10.5281/zenodo.6426938.

Contribuciones	Firma
Investigación previa	
Redacción de las respuestas	
Desarrollo del código	

Bibliografía

1. **Subirats Maté, Laia y Calvo González, Mireia.** *Web scraping.* PID_00256970.

2. **Mitchell, Ryan.** *Web Scraping with Python.* s.l. : O´Reilly, 2018. ISBN 978-1-491-98557-1.