Práctica 1: Web Scraping

Alumna: Helene Jordan Parize

Aula 2

Tipología y Ciclo de Vida de los Datos

Curso 2019-2020

Universitat Oberta de Catalunya

Helene Jordan Parize

Contexto

Para el estudio de la evolución de la música a lo largo de los años, es interesante explorar las tendencias musicales más populares a lo largo de los años. Para muchas radios o plataformas de streaming como Spotify o SoundCloud es importante saber cuáles son las canciones más escuchadas y qué características tienen para poder crear listas de reproducción más universales, con más seguidores a la vez que ver cómo se pueden relacionar estas canciones con los recomendadores más personalizados.

Un paso para este estudio, es ver cuáles son las canciones más escuchadas (por plataformas streaming, o por la radio) y vendidas (tanto física como digitalmente). La lista de Hot 100 de Billboard, recopila esta información desde 1958 en Estados Unidos y es una de las listas más importantes a nivel musical que se actualiza cada semana y se puede ver la evolución de las canciones dentro de ella.

La práctica de web scraping realizada, se encarga pues de recolectar la información presente en la web de billboard.com, en particular, en la lista Hot 100.

Dataset

Título: **yearly_hot_100_chart**

Descripción: Este dataset recoge la información de la lista de las canciones más escuchadas semanalmente durante un año determinado en Estados Unidos. Algunos de los datos que se recogen son: el título, el nombre del artista, la posición que ocupa la canción en la lista y la semana.

Helene Jordan Parize

Imagen

Contenido

Este dataset contiene datos anuales. Billboard lleva desde el 4 de agosto de 1958 la lista de canciones más vendidas y escuchadas en Estados Unidos semanalmente. De esta manera, el código se puede ejecutar para todos los años entre 1958 hasta el actual.

Los atributos recogidos son:

- **artista**: con el nombre de/los artista/s que interpretan la canción.
- **last_week**: qué posición ocupaba la canción la semana pasada (en caso de no estar en la lista la semana pasada se marca '-') .
- **name**: nombre de la canción
- **peak**: posición más alta en la que ha estado la canción a lo largo de toda su historia.
- **rank**: posición que ocupa en la lista la canción durante la semana
- **rise**: variable de 4 factores. ¿La canción ha aumentado o ha disminuido de posición? ¿Es nueva o ha vuelto a aparecer en la lista?
- **rise_nb**: si la canción ha aumentado o disminuido de posición de la semana pasada a la actual, cuántas posiciones se ha desplazado. La diferencia entre aumentar y disminuir se marca con el signo del dígito (negativo o positivo). Si no se ha desplazado se marca con '-'.
- **week**: la semana de la cual se ha extraído la lista.
- **week_chart**: cuántas semanas ha estado la canción en la lista.

El web scraping se ha realizado con el lenguaje Python gracias a las librerías requests y BeautifulSoup.

Agradecimientos

Los datos han sido recolectados desde la página web de https://www.billboard.com/.

Inspiración

El dataset recogido en esta práctica puede ser utilizado en muchos contextos. Una de los ejemplos de estudio de los cuales se podría tomar inspiración es este vídeo: https://www.youtube.com/watch?v=qJT2h5uGAC0.

En este vídeo, se explora la tendencia en la historia de la música del éxito de las canciones interpretadas por un artista masculino cantando algunas partes en *falsetto*. En el video, uno de los conjuntos de datos que sirven para realizar el análisis son las listas hot 100 de Billboard a lo largo de los años. Este conjunto de datos se combina con las características de las canciones y la presencia de *falsetto*.

Por lo tanto, uno de los objetivos que podría tratar de resolver este conjunto de datos sería averiguar alguna tendencia a lo largo de la historia de la música como, por ejemplo, ¿qué es lo que hace que una canción sea un éxito? En este caso, el conjunto de datos se debería combinar con otros datos que describiesen características de las canciones para poder elaborar un modelo de minería de datos que generase un árbol de decisión o utilizase una red neuronal para determinar qué características hacen que una canción sea un éxito.

Por otro lado, también puede servir para que plataformas de streaming como Spotify cree listas de reproducción de canciones exitosas de décadas pasadas. Es decir, es cierto que Spotify, por ahora tiene datos del número de streams que tiene cada canción, y por lo tanto, es más fácil generar listas de reproducción de las canciones más escuchadas en la década de 2010. Sin embargo, no contiene datos de las canciones más escuchadas en 1970 puesto que la plataforma todavía no existía. Gracias al conjunto de datos, se podría crear una lista de reproducción con las canciones más escuchadas por entonces.

Licencia

La licencia escogida para el conjunto de datos presentado en este trabajo es: CC BY-SA 4.0. De esta manera se aseguran las condiciones siguientes:

- Se debe proveer el nombre de la creadora del trabajo. De esta manera, se reconoce el trabajo y la aportación de la autora.

- Se puede hacer uso comercial de los datos. De esta manera, otras empresas o fuentes de periodismo pueden hacer uso del conjunto de datos.
- Share-alike, así las contribuciones posteriores a este se basan en licencias iguales o más restrictivas que esta. De esta manera, se asegura que las condiciones de la licencia actual se sigan manteniendo.