Curso: Data Science
Comisión: 39960
Fecha de entrega: 26 de abril de 2023
Nombre: Catalina Miranda

Desafío N°8: Descarga de datos desde APIs públicas

ABSTRACT

La temática que se desarrollará en el presente proyecto final se relaciona con el mundo de la música, más específicamente, con el ranking de la mundialmente reconocida revista Billboard. Desde el año 1958, cada semana, la institución da a conocer los "Hot 100", es decir, las 100 canciones más aclamadas del momento. La clasificación no solo está basada en las ventas físicas y digitales de los Estados Unidos, sino también en las reproducciones por radio y por plataformas de *streaming* a nivel global.

La base de datos seleccionada agrupa información relacionada con las canciones y los artistas que lograron ingresar al Hot 100 del ranking de la revista desde el año 1958 hasta el 2021. La información presente en la base de datos se complementará con información extraída desde un API diseñado específicamente para extraer Billboard charts desde Python. Luego de este proceso, se obtendrá información del hot100 desde 1958 hasta 2022 inclusive. Por cada semana, se muestran la posición y los títulos de las 100 canciones de la clasificación junto con sus correspondientes los cantautores o grupos musicales. Además, el conjunto de datos incluye información técnica de las canciones presentes en el ranking, como la valencia, el tempo, el modo y la acusticidad, entre otros.

A partir de la información histórica, se identificarán los artistas con mayor trayectoria y las canciones más exitosas. Por otro lado, en base a los datos técnicos de las canciones del ranking, se podrá establecer si existen patrones musicales que hacen que una canción sea más escuchada que otra. En un contexto comercial, dicha información resulta de interés para los productores musicales y compañías discográficas, porque les permite conocer las características que deberían tener las nuevas melodías para convertirse en "*hits*" mundiales.

PROBLEMA Y CONTEXTO COMERCIAL

Una discográfica está teniendo problemas con los artistas que contrata. Las canciones que producen no llegan a posicionarse dentro de las canciones más escuchadas, y las recaudaciones de la compañía, junto con su reputación, están cayendo drásticamente. Mientras tanto, otras empresas del ámbito musical están creciendo notablemente. Por tal motivo, la discográfica desea conocer cuáles son las características de las canciones más exitosas, para poder predecir cuáles van a ser parte del "Hot 100" y cuáles no. El objetivo principal es implementar un modelo de *Machine Learning* que indique las probabilidades de que la canción se posicione entre las mejores del mundo.

RECOLECCIÓN DE DATOS

El conjunto de datos está compuesto por dos archivos en formato csv: el primero llamado "Top 100" que contiene 12 columnas y 327896 filas, y el segundo denominado "Audio" que posee 19 columnas y 29504 registros. Ambos archivos están relacionados por el campo SongID, que permite identificar de manera unívoca a cada una de las canciones que aparece en el set de datos.

A continuación, se especifican los atributos de cada archivo.

"Top 100" indica las canciones que ingresaron al top 100 del Billboard Chart entre 1958 y 2021. Los campos son los siguientes:

1. Url: link que permite ingresar al sitio web de Billboard y visualizar el ranking de la semana correspondiente.
2. Week position: posición de la canción en el top 100 de la semana correspondiente
3. Song: nombre de la canción
4. Performer: cantante/grupo musical
5. SongID: clave primara que relaciona las tablas
6. Instance: indica la cantidad de veces que ha aparecido la canción en el top 100
7. Previous week position: indica la posición en el ranking la semana anterior a la analizada
8. Peak position: indica la máxima posición alcanzada por la canción en el ranking.
9. Weeks on chart: indica la cantidad de semanas de la canción en el ranking.
10. Day: día de la semana en la que se publicó el ranking
11. Month: mes en el que se publicó el ranking
12. Year: año correspondiente al que se publicó el ranking

"Audio" especifica los atributos musicales de las canciones que ingresaron al ranking.

1. SongID: clave primara que relaciona las tablas
2. Spotify_genre: género según la clasificación de Spotify

3. Spotify_track_ID: código que identifica de forma unívoca a la canción en Spotify
4. Spotify_track_preview_url: link que permite acceder a 30 segundos de la canción.
5. Spotify_track_duration_ms: duración de la canción en milisegundos.
6. Spotify_track_explicit: "*Explicit*" es una clasificación de Spotify que significa que la canción usa lenguaje grosero, violento u ofensivo. Se indica como verdadero o falso.
7. Spotify_track_album: álbum al que pertenece la canción.
8. Danceability: describe qué tan bailable es la canción, en función de elementos como el tempo y la estabilidad del ritmo. 0 significa que es muy poco bailable y 1 significa que es 100% bailable.
9. Energy: es una medida que varía entre 0 y 1 y que representa un porcentaje de intensidad de actividad. Normalmente, una canción enérgica se suele rápida, fuerte y ruidosa.
10. Key: es el tono, las notas o la escala de la canción que forma la base. Las 12 claves varían entre 0 y 11.
11. Loudness: es la calidad de la canción, desde -60 a 0 dB promediada a lo largo de toda su duración. Cuanto mayor es el valor, más alta es la canción.
12. Mode: las canciones se clasifican en menores (0) o mayores (1).
13. Speechiness: detecta la presencia de palabras en la canción. 1 significa que el 100% del tiempo se dicen palabras.
14. Acousticness: o acusticidad, que es la propiedad de lo que suena agradablemente, 1 significa que la canción es acústica, 0 significa que no es nada acústica.
15. Instrumentalness: indica si existe contenido vocal. 1 significa que no hay contenido vocal. La saqué
16. Liveness: detecta la presencia de audiencia durante el grabado de la canción. Cuanto más cerca de uno se encuentra el valor, mayor probabilidad de que se haya grabado en vivo.
17. Valence: es una medida que varía entre 0 y 1, y que describe la positividad musical transmitida por una pista. Las pistas con una valencia alta suenan más positivas (felices, alegres, eufóricas, etc), mientras que las pistas con una valencia baja suenan más negativas (tristes, deprimidas, enfadadas, etc).
18. Tempo: es la velocidad de la canción y se mide en BMP (*Beats por minuto*).
19. Spotify_track_popularity: mide la popularidad de la canción en Spotify. Los valores varían entre 0 y 100, indicando qué tan popular es el artista en relación a todos los demás artistas de la plataforma.

Link de descarga del conjunto de datos seleccionado:

https://www.kaggle.com/datasets/heemalichaudhari/billboard-hot-weekly-charts

Al conjunto de datos seleccionado se le anexará información descargada desde un API.

En primer lugar, se descargarán los Billboard charts faltantes del año 2021 y se agregará al análisis el año 2022 completo. Para ello, se usó el API billboard.py, disponible en Python para acceder a los datos del ranking Hot100. Luego de su instalación se especificaron las semanas del ranking deseadas. Dado que se trata de una cantidad interesante de semanas, se generó una serie con la fecha inicial (mayo de 2021), y cada una de las fechas posteriores se calcularon sumando 7 días. Los datos obtenidos se agregaron a los archivos planos, para luego poder ser anexados al modelo relacional planteado para el presente proyecto final.

Por otro lado, al extender la cantidad de cuadros analizados, se encontraron nuevas canciones que no poseen datos de sus atributos musicales. Para completar la tabla Audio, y minimizar la cantidad de datos ausentes, los atributos de las canciones se buscaron mediante el API web de Spotify, disponible para todos sus usuarios. Este proceso también fue útil para completar los datos que originalmente faltaban en el archivo Audio.