Documentación del Dataset songs.csv

1. Descripción general

El dataset songs.csv contiene información detallada sobre canciones y artistas, con un total de 39 columnas. Cada fila representa una canción única, y el archivo está ordenado alfabéticamente por el nombre del artista (Artist(s)), lo que permite realizar búsquedas secuenciales eficientes por artista una vez identificado su offset inicial.

Este dataset se utiliza como base para realizar búsquedas de canciones a través de un sistema cliente-servidor (implementado en C), donde el servidor recorre el archivo desde los offsets precalculados para cada artista.

2. Estructura del dataset

Campo	Descripción	Tipo / Ejemplo
bucket	Índice hash (0–349) que identifica el bucket del artista (MD5 mod 350).	Entero
Artist(s)	Nombre del artista o grupo principal (clave principal, orden alfabético).	Texto
song	Título de la canción.	Texto
text	Letra o fragmento representativo de la canción.	Texto largo
Length	Duración de la canción.	Texto, '3:45'
emotion	Emoción predominante.	Texto
Genre	Género musical.	Texto
Album	Nombre del álbum.	Texto
Release Date	Fecha de lanzamiento.	Fecha
Key	Tonalidad o clave musical.	Texto
Tempo	Tempo en BPM.	Numérico
Loudness (db)	Volumen promedio en decibelios.	Numérico

Time signature	Compás musical.	Texto o entero
Explicit	Lenguaje explícito (True/False).	Booleano
Popularity	Puntuación de popularidad (0–100).	Numérico
Energy	Nivel de energía (0–1).	Decimal
Danceability	Qué tan bailable es la canción (0–1).	Decimal
Positiveness	Nivel de positividad (0–1).	Decimal
Speechiness	Porcentaje de palabras habladas (0–1).	Decimal
Liveness	Medida de interpretación en vivo (0–1).	Decimal
Acousticness	Nivel de acústica (0–1).	Decimal
Instrumentalness	Proporción instrumental (0–1).	Decimal
Good for Party	Adecuada para fiestas.	Booleano
Good for Work/Study	Adecuada para trabajo o estudio.	Booleano
Good for Relaxation/Meditation	Adecuada para relajación o meditación.	Booleano
Good for Exercise	Adecuada para ejercicio.	Booleano
Good for Running	Adecuada para correr.	Booleano
Good for Yoga/Stretching	Adecuada para yoga o estiramiento.	Booleano
Good for Driving	Adecuada para conducir.	Booleano
Good for Social Gatherings	Adecuada para reuniones sociales.	Booleano
Good for Morning Routine	Adecuada para la rutina matutina.	Booleano

Similar Artist 1–3	Artistas similares.	Texto
Similar Song 1–3	Canciones similares asociadas.	Texto
Similarity Score 1–3	Puntaje de similitud (0–1).	Decimal
h_artist	Hash MD5 hexadecimal del artista.	Texto

3. Características relevantes

• Ordenado alfabéticamente por artista, lo que permite cortar la búsqueda secuencial cuando se detecta un cambio de artista.
• Campo Bucket obtenido del hash MD5 version corta y haciendo mod 350 del nombre del artista, columna creada por nuestro grupo, en Python.
• h_artist es el hash MD5 completo en hexadecimal.
• index.csv almacena los offsets de la primera aparición de cada artista.
• El archivo tiene alrededor de 500k filas, y pesa alrededor de 1GB

4. Uso en nuestro programa:

El dataset predeterminado ya esta de ordenado alfabéticamente por artistas, decidimos hacer este nuestro campo principal, hay alrededor de 120k artistas únicos, y estos artistas, los distribuimos en Buckets usando una version corta de hash MD5 y posteriormente sacándole el módulo 350 al hash, con esto tenemos 350 Bucket con alrededor de 350 objetos cada uno, así, como máximo, se tendrían que hacer 350 búsquedas de buckets más 350 búsquedas dentro del bucket, dando como máximo 700 búsquedas. Todo esto aprovechando que debajo de la primera canción del artista, están todas las demás canciones del mismo artista.

El programa tiene un funcionamiento simple, el usuario ingresa el artista y el nombre de la canción, con estos datos, el programa calcula el bucket de dicho artista, ahí, empieza a recorrer linealmente dicho bucket buscando el artista, cuando lo encuentra, empieza a iterar sus canciones, hasta encontrar la canción o terminar todas las canciones del artista