YouTube Playlist Scraper for Catalan Music

This script downloads **MP3 audio tracks** from a list of **YouTube playlists** and saves structured **metadata** for each song. It is intended for building datasets of Catalan music for analysis, classification, or archival purposes.

🧠 What It Does

- Loads YouTube playlist URLs from a CSV file.
- For each video:
    - Extracts metadata (title, uploader, video ID).
    - Downloads the audio in high-quality MP3 format.
    - Cleans and standardizes the song title.
    - Records metadata in a CSV file (`catalan_music_metadata.csv`).
- Skips videos that have already been downloaded (tracked by video ID).

📦 Dependencies

Install the required Python packages:

```
pip install yt-dlp pandas
```

Also ensure that `ffmpeg` is installed and available in your system path (used for audio conversion by yt-dlp).

📁 Input File Format

Create a file named `playlists.csv` in the same directory as the script. It must have one column called `playlist_url`, like this:

```
playlist_url
https://www.youtube.com/playlist?list=PLabc123...
https://www.youtube.com/playlist?list=PLxyz456...
```

Each row should be a full YouTube playlist URL.

🚀 How to Run

Place the script in your working directory along with the `playlists.csv` file, then run:

```
python youtube_playlist_scraper.py
```

This will:

- Download MP3 audio files to the `downloads/` folder.
- Create or update `catalan_music_metadata.csv` with metadata for each downloaded song.

📄 Output

🎵 Audio Files

Stored in the `downloads/` folder as:

```
videoid_cleaned_song_title.mp3
```

-

📝 Metadata File

- `catalan_music_metadata.csv` contains:
    - `video_id`: YouTube video ID
    - `song_title`: Cleaned and normalized song title
    - `channel_name`: Uploader/channel name
    - `original_title`: Original YouTube video title
    - `audio_file`: Filename of the downloaded MP3

Example:

```
video_id,song_title,channel_name,original_title,audio_file
abc123,La_Rumba_Bonita,CatalanSounds,La Rumba Bonita - CatalanSounds,abc123_La_Rumba_Bonita.mp3
```

🧽 Title Cleaning Logic

- Removes channel name from video title.
- Eliminates special characters (`\, /, :, *, ?, etc.`) and fancy quotes.
- Replaces spaces with underscores.
- Produces safe and consistent filenames across systems.

🔄 Redundancy Handling

The script checks for previously downloaded video IDs to avoid re-downloading the same tracks. You can delete entries from `catalan_music_metadata.csv` to reprocess them.

💡 Notes

- Script handles playlist parsing errors gracefully and continues processing others.
- Only public playlists and videos can be accessed.
- Downloads the **best available audio** and converts it to 192 kbps MP3.

📬 Contact

For dataset structure or further automation, contact the developers of the MTG-102 project or the Music Technology Lab team.