Introducció

Estructurarem el contingut d'aquesta carpeta en dues parts. Per una banda s'han computat els èxits Spotify, Grammy i Billboard d'un subconjunt de 55.000 cançons del MillionSongsDataset. Per altra banda, s'han estudiat les cançons dels 31 artistes del MillionSongsDataset amb el valor "artist_hotness" més elevat i s'han calculat els tres èxits d'aquestes cançons.
Per a extreure la popularity de les cançons s'han utilitzat els scripts playlist_creator.py i data_scraping.py, penjats en format .py. Per al tractament de dades s'han utilitzat notebooks. Aquests s'han penjat en format .pdf per a que es puguin veure els outputs. Tant els .py com els notebooks es troben en aquesta mateixa carpeta.
Els .csv amb els que hem treballat es troben emmagatzemats en subcarpetes que anirem detallant en aquest README per a una millor organització de les dades.

Primera part: Subconjunt de 55.000 cançons

1. Creació del dataset amb la popularitat de Spotify

- **playlist_creator.py:** itera per les cançons del **msd_reduced.csv** (es troba a la carpeta reduced_datasets), genera de manera iterativa llistes de reproducció de 1000 cançons trobades a Spotify i les va afegint a un compte de Spotify Developer.
- **data_scraping.py:** itera per les llistes de reproducció generades i emmagatzema en .csv separats per cada llista de reproducció el títol, artista i popularity de la cançó. Els fitxers resultants es poden consultar a la carpeta **spotify_subsets.**
- Generem així la carpeta de csv **spotify_subsets**, que conté {title, artists, popularity} de les cançons trobades per Spotify dins el subset de 55000 cançons del MillionSongsDataset.
- Amb el notebook **combine_spotify_csv** ajuntem tots els datasets de **spotify_subsets** en un de sol, anomenat **spotify_combined.csv**. Netegem el dataset eliminant les files duplicades. Considerem files duplicades aquelles amb mateixa clau primària: mateix titol i artista. El dataset **spotify_combined.csv** resultant es pot trobar a la carpeta **reduced_datasets.**

2. Neteja de les dades en els datasets MillionSongs, Spotify i Grammy

- Amb el notebook **data_cleaning** netegem els datasets **million_songs_dataset.csv**, **billboard.csv** i **grammy.csv** segons els paràmetres estipulats a la memòria. Els datasets sense modificar, els originals, es troben a la carpeta **original_datasets**. Obtenim els datasets reduïts **msd_reduced.csv**, **billboard_reduced.csv** i **grammy_reduced.csv**. Aquests es troben a la carpeta **reduced_datasets**.

3. Càlcul dels tres èxits per al subconjunt de 55.000 cançons del MillionSongs

- El notebook **success_subset_55000_spotify** calcula les tres mesures d'èxit plantejades a la memòria per al subconjunt de 55000 primeres cançons del Million Songs Dataset. Treballa amb els datasets **msd_reduced.csv**, **billboard_reduced.csv**, **grammy_reduced.csv** i **spotify_combined.csv**, que es troben a la carpeta **reduced_datasets**. Genera un nou dataset amb el subconjunt

del conjunt de les 55000 primeres cançons que es troben a Spotify. Conté el títol, artistes, les mesures empíriques extretes del msd d'aquestes cançons i els valors dels tres èxits: e_spotify, e_grammy i e_billboard. El dataset resultant s'anomena **successes55000_with_spotify.csv.** Es troba a la carpeta **datasets_with_successes.**

Segona part: 31 artistes més populars

1. Creació del dataset amb la popularitat de Spotify

- Amb el notebook **spotify_songs_by_artists** generem un subconjunt del MillionSongsDataset que contingui les cançons dels 31 artistes amb un artists_hotness més elevat. Aquests són: "Johnny Cash", "The Rolling Stones", "Bruce Springsteen", "Ray Charles", "Muse", "Céline Dion", "Michael Jackson", "U2", "Green Day", "Jason Mraz", "Adam Sandler", "Michael Bublé", "Mariah Carey", "Linkin Park", "Brad Paisley", "Queen", "T.I.", "Ustad Nusrat Fateh Ali Khan", "Metallica", "Daft Punk", "Eminem", "Coldplay", "James Brown", "The Black Keys", "Enrique Iglesias", "John Frizzell", "Harry Connick_ Jr.", "Weezer", "Nickelback", "Mannheim Steamroller", "DJ Bobo".
- Aquest nou csv s'anomenarà **artists_hotness.csv**. Es troba a a carpeta **artists_hotness_lists**.
- **playlist_creator.py:** itera per les cançons del **artists_hotness.csv**, genera de manera iterativa llistes de reproducció de 1000 cançons trobades a Spotify i les va afegint a un compte de Spotify Developer.
- **data_scraping.py:** itera per les llistes de reproducció generades i emmagatzema en .csv separats per cada llista de reproducció el títol, artista i popularity de la cançó. Els fitxers resultants es poden consultar a la carpeta **spotify_subsets_artists.**
- Generem així la carpeta de csv **spotify_subsets_artists**, que conté {title, artists, popularity} de les cançons trobades per Spotify dins el subset de les cançons dels artistes més populars **artists_hotness.csv**.
- Amb el notebook **combine_spotify_csv** ajuntem tots els datasets de **spotify_artists** en un de sol, anomenat **spotify_combined_artists.csv**. Netegem el dataset eliminant les files duplicades, amb mateix titol i artista. Es troba a **reduced_datasets**.

2. Càlcul dels tres èxits per al subconjunt de les cançons dels artistes amb major artists_hotness

- El notebook **success_with_songs_artists** calcula les tres mesures d'èxit plantejades a la memòria per al subconjunt de cançons dels 31 artistes amb l'atribut artist_hotness més elevat. Treballa amb els datasets **msd_reduced.csv**, **billboard_reduced.csv**, **grammy_reduced.csv** i **spotify_combined_artists.csv**, que es troben a la carpeta **reduced_datasets**. Genera un nou dataset amb el subconjunt del conjunt de les cançons de **artists_hotness.csv**. Conté el títol, artistes, les mesures empíriques extretes del msd d'aquestes cançons i els valors dels tres èxits: e_spotify, e_grammy i e_billboard. El dataset resultant s'anomena **successes_with_artists.csv.** Es troba a la carpeta **datasets_with_successes.**