MILESTONE 1 - Data Visualization (COM-480)

AFFOLTER Joanne - 353166
GOMEZ DONOSO Damien - 296644
ELTAYAR Esraa - 343960

Link to the google docs (to access the urls) :
https://docs.google.com/document/d/10ix4ZB86Yu50BtoEO-9mhvX9OvTdcPExjZ1oRRP6crM/edit?usp=sharing

TITLE

How do we link a song's mood to our emotions, and are these connections valid?

DATASET

Playlists

URL : https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge
Description : The dataset contains 1,000,000 playlists, including **playlist titles** and **track titles**, created by users on the Spotify platform between January 2010 and October 2017.
Preprocessing : The dataset is in JSON format, so we will only retrieve the name and id of the playlists and the tracks contained in these playlists. We will create a file in .csv format, in the form (playlist_id, playlist_name, track_id, track_name). This will allow us to easily identify which playlists the tracks belong to and which tracks appear together frequently in Spotify users' playlists.

Tags

URL : MuSe: The Musical Sentiment Dataset | Kaggle
Description : The MuSe dataset contains sentiment information for 90,001 songs. Each track is associated with **social tags**, which have seeded the scraping of the latter on Last.fm.
Preprocessing : We will base our study on the tracks of this dataset. We only keep the track name, artist, genre, tags and spotify_id. This last one will allow us to complete the dataset with the audio features of the tracks retrieved from the Spotify API.

Emotions

URL : https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
Description : Manually generated dataset containing a list of **words** and their associations with eight basic **emotions** : disgust, joy, sadness, surprise, trust, anticipation, anger, fear

plus two sentiments, positive and negative.

Preprocessing : We convert the file into a dataframe and remove the rows where all basic emotions are set to zero. Words that do not appear in the tags studied are also removed. This dataframe will be used to associate emotions with each track in the ***Tags*** dataset based on their tags.

PROBLEMATIC

Our project aims to visualise the fascinating link between music and emotion. Specifically, we plan to explore the relationship between different groups of tracks characterised by the same emotion. Do the tracks in the same playlist elicit similar emotions? Can we identify any trends in the association of emotions among playlists? If so, what are these trends? Additionally, we aim to investigate whether people create playlists randomly or if they are mindful of the emotions they wish to evoke through their song selection.

Another key focus of our project is to cluster tracks based on emotions, drawing on the work of Robert Plutchik (https://en.wikipedia.org/wiki/Robert_Plutchik) . We intend to classify our tracks into eight basic emotions (disgust, joy, sadness, surprise, trust, anticipation, anger, and fear) and eight mixed emotions (love, submission, awe, disapproval, remorse, contempt, aggressiveness, and optimism). Our goal is to analyse the composition of these clusters, exploring the artists represented, the distribution of genres across emotions, and the potential presence of common musical features. Furthermore, we seek to identify any predominant emotions within the tracks and evaluate their overall balance.

As music lovers ourselves, we were naturally drawn to this project, inspired by the intriguing connection between music and emotion. Moreover, we believe that this project is accessible to everyone, as one need not possess deep musical knowledge to appreciate the emotional impact of music.

EXPLORATORY DATA ANALYSIS

PLAYLISTS :

After building a dataframe with columns (playlist_id, playlist_name, track_id, track_name) from the JSON file, we created the following graph to model the associations between tracks in Spotify playlists:

- Nodes: tracks
- Weighted edge (u,v) ⇔ tracks u and v appear in same playlist (weight = # playlists they appear together)

TAGS :

We keep the tracks that have a Spotify ID and fetch their audio features through the Spotify API to enrich our data. Since a track can have multiple tags, we use Pandas' pd.explode() to create multiple rows for each track, each containing a single tag. We then determine the number of occurrences of each tag in the whole dataset and keep only those that appear more than 50 times. Tracks that no longer have associated tags are removed. We also store all the (unique) tags in a set for future use.

EMOTIONS :

After selecting the words of interest ("tags" of the tracks studied), we associate them with an "emotion vector" where each binary component characterises one of the 8 basic emotions: 1 if the word conveys the emotion, 0 otherwise.

We merge these vectors to the Tracks dataset based on the tag attribute. Each track is therefore associated with one or more "emotion vectors", which we sum together to obtain a single vector v.

We then simplify the vectors to make them similar for several tracks.

To do this:

– We create a binary vector $w,\ s.t.\ wi\ =\ 1\quad \text{if}\ i\ =\ argmax_k v$,

$$0\quad \text{otherwise}$$

– We retain only the simplified vectors that correspond to the emotions targeted by Plutchik (basic and mixed emotions)

Finally, we group the data by their emotion vector to obtain clusters of tracks according to their principal sentiment. The final dataset thus contains an "emotions" column that associates an emotion to each track.

This figure displays the final clusters of emotions that we were able to identify. Each mixed and basic emotion is represented, along with the number of tracks associated, and their corresponding IDs.

	new_key	id	counts	emotions
0	00000001	[4710, 14603, 19756, 2940, 4067, 5160, 9431, 9...	40	anticipation
1	00000010	[1752, 1753, 1755, 1756, 1757, 1759, 1760, 176...	512	anger
2	00000100	[4099, 4889, 4955, 5088, 5182, 8907, 11562, 11...	149	disgust
3	00000110	[820, 821, 822, 823, 824, 825, 826, 828, 829, ...	751	contempt
4	00001000	[1638, 1698, 1699, 1701, 1702, 1703, 1706, 174...	3433	sadness
6	00001100	[1647, 2784, 6467, 6531, 8249, 8510, 8897, 891...	708	remorse
8	00010000	[9429, 9440, 9466, 22500]	4	surprise
10	00100000	[4036, 5212, 6309, 6311, 6342, 6346, 7709, 776...	834	fear
24	00110000	[3271, 3343, 3443, 3488, 3737, 3938, 3946, 3485]	8	awe
35	01000000	[4069, 4972, 4974, 5355, 6428, 6614, 8846, 114...	705	trust
41	01100000	[20465, 20524]	2	submission
47	10000000	[3001, 4039, 4058, 4059, 4081, 4122, 4721, 476...	3516	joy
48	10000001	[5214, 7738, 12654, 12753, 12897, 12909, 12952...	878	optimism
85	11000000	[1714, 4728, 4732, 4891, 4914, 4952, 4987, 507...	1450	love

RELATED WORK

The paper related to the dataset ***Playlist*** is [Recsys challenge 2018 | Proceedings of the 12th ACM Conference on Recommender Systems](). We can find this dataset on AIcrowd. This dataset was used for a challenge called ***Spotify One Milion Playlist Dataset Challenge.*** The purpose of this challenge was to use this dataset in order to develop a system for the task of automatic playlist continuation. More especially, the system should choose new tracks to add into a given dataset according to its composition.

The ***Emotions*** dataset is based on the two works done here https://saifmohammad.com/WebDocs/Mohammad-Turney-NAACL1EmotionWorkshop.pdf and here https://arxiv.org/pdf/1308.6297.pdf. This dataset is used for applications like works on computational literature. For example, the work From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales use this dataset for sentiment analysis into both individual books and very large collections of books.

The work that explains the creation of the ***Tags*** dataset is explained here Toward a Musical Sentiment (MuSe) Dataset for Affective Distant Hearing.

What sets our project apart is our unique approach, which combines data from three separate datasets and merges their information. Additionally, we have added a new feature to the tracks, inspired by Robert Plutchik's theory of emotions, providing a novel perspective on music and its emotional impact.

Moreover, we aim to integrate two distinct types of visualisations into one cohesive output, making our project particularly intriguing. Our project draws inspiration from notable visualisations, such as Plutchik's Wheel of Emotions: Feelings Wheel • Six Seconds and Hierarchical edge bundling – from Data to Viz allowing us to explore music from multiple angles and generate fresh insights into the emotional associations with music.