134 Final Project: Music Recommender via Content-Based Recommendations, Matrix Factorization and Web-Scraping

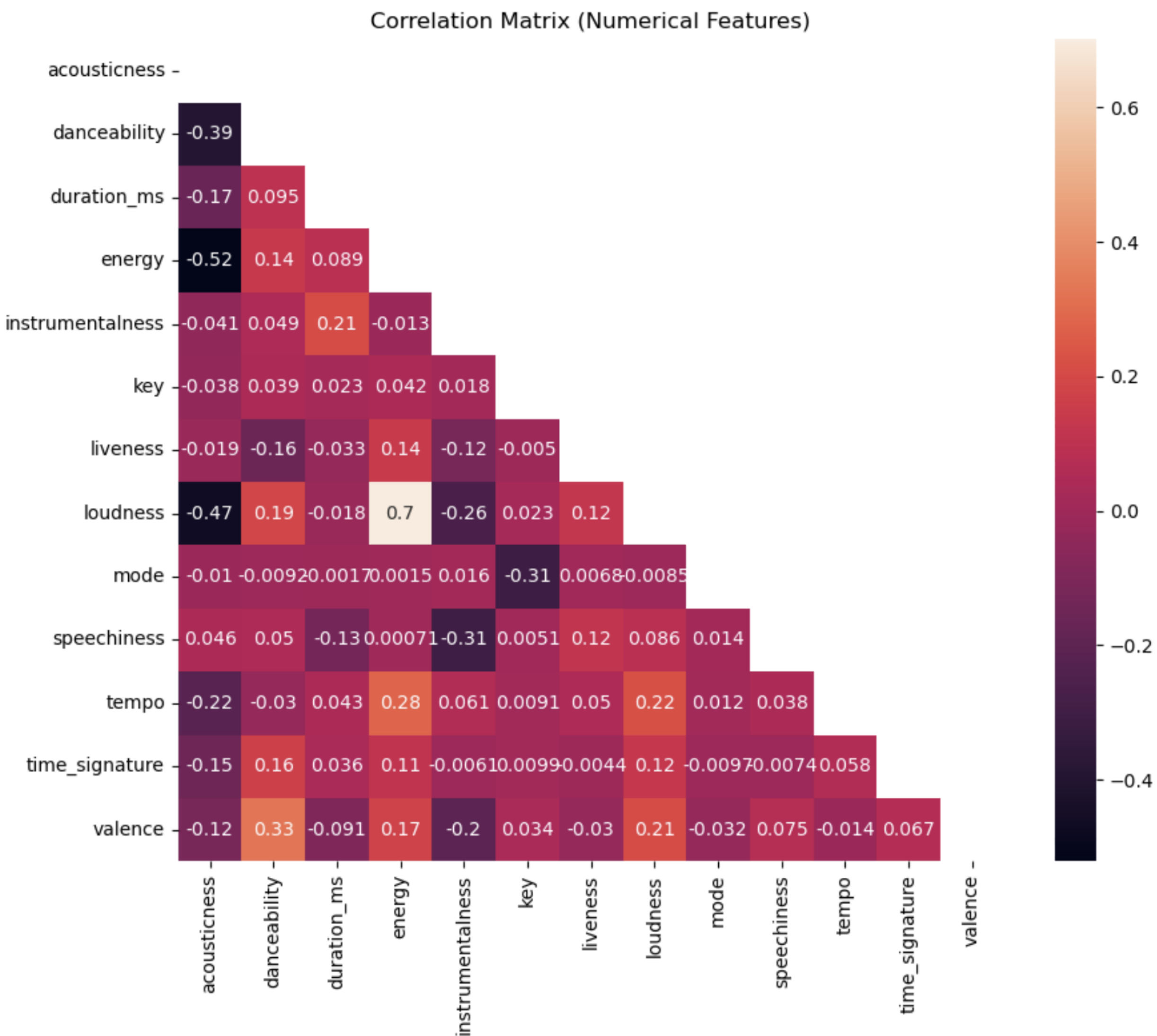
Group 6: Arthur Kim, Brian Sun, Kira Jackson, Meghana Dhruv

2024-12-12

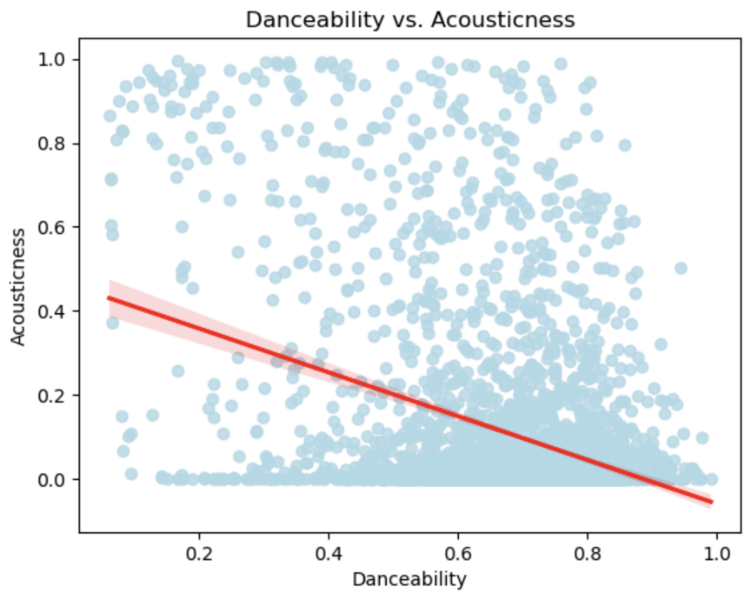
Introduction & EDA

The purpose of this project is to develop a content-based and matrix factorization music recommendation system by leveraging song metadata and simulated user data. This system aims to identify and recommend songs that share similar characteristics, enhancing the user's experience through personalized suggestions. The foundation of this project lies in a Kaggle dataset, *"10+ M. Beatport Tracks / Spotify Audio"* which provides comprehensive audio feature data for a large collection of songs in the audio_features.csv file. These features include attributes like danceability, energy, valence, tempo, acousticness, and liveness. A key limitation of this dataset was the lack of artist names and song titles, which are critical for presenting meaningful and user-friendly recommendations. Instead, the dataset included ISRC (International Standard Recording Code) values as unique identifiers for songs. Thus, our group developed a web scraper to retrieve the song titles and artist names by scraping *soundexchange.com* using the ISRC values. Using this enriched dataset, we built two types of recommender systems: content based and matrix factorization.

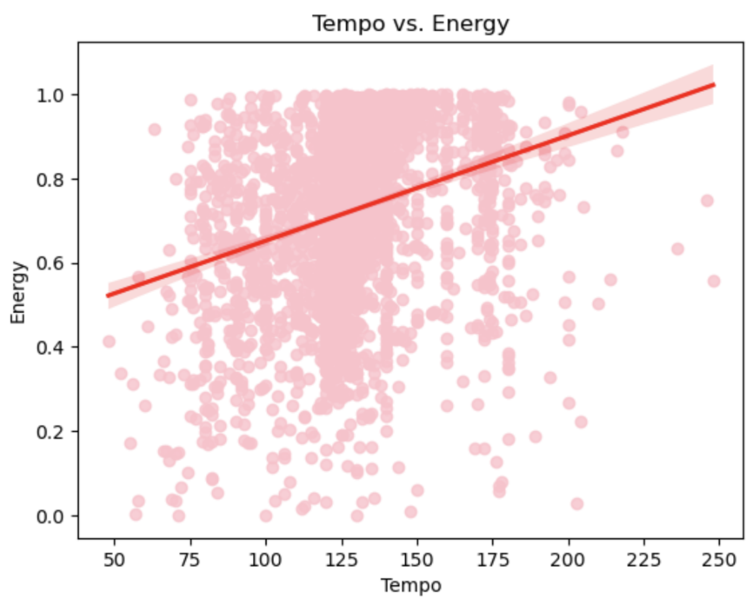
Exploratory Data Analysis (EDA) was performed to understand the relationships between various audio features in the dataset and their potential impact on the recommendation system. This analysis provided insights into the correlations and distributions of features.



The correlation matrix above highlights the relationships between numerical audio features in the dataset. A notable positive correlation is observed between loudness and energy 0.7, indicating that tracks with higher loudness levels tend to have higher energy. Acousticness is negatively correlated with energy -0.52 and loudness -0.47. This indicates that more acoustic tracks tend to be quieter and less energetic, typically found in softer, instrumental genres.



This scatter plot explores the relationship between danceability and acousticness. A clear negative trend is observed, with higher danceability associated with lower acousticness. Tracks with high danceability are generally less acoustic which aligns with electronic or pop genres.

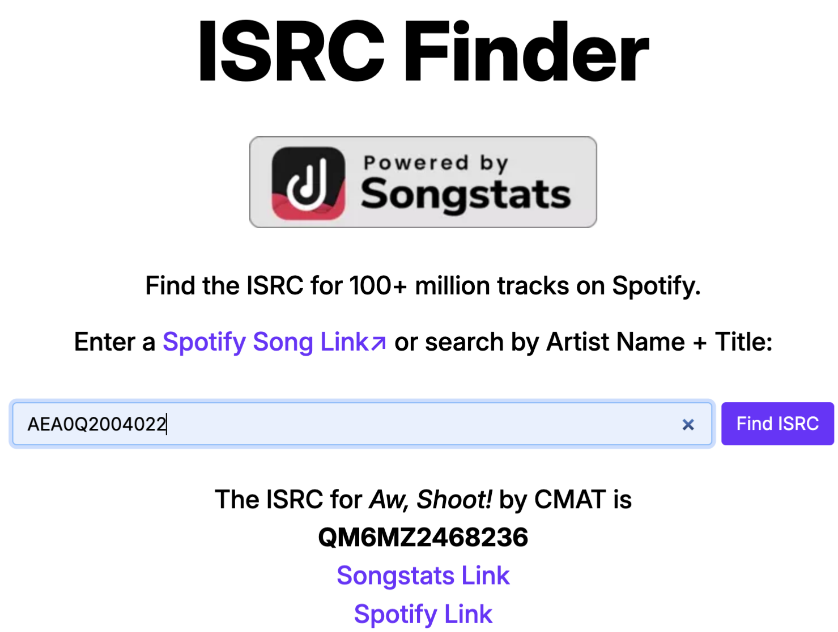


This scatter plot examines the relationship between tempo and energy. This positive trend indicates that tempo is a key feature for categorizing high-energy songs. The insights gained from the plots helped refine the selection of features used in our recommendation model.

Methods/Results

Web-Scraper

Using Python, the audio_features.csv file from Kaggle was read in, the primary key being the ISRC codes. The Selenium library was used to automate web browser interactions and for each ISRC code, a URL was constructed to search for the rest of the song information on soundexchange.com. Once the page loaded, the algorithm located the table elements containing the search results. ThreadPoolExecutor was used to execute the web scraping for multiple ISRC codes to optimize the recommender even further. To prevent overwhelming the target website, a throttle was implemented by introducing delays between requests and adding randomized intervals to mimic human behavior, thereby reducing the risk of being detected or blocked. Up to 6 ISRC codes could be run at once and if an invalid ISRC was inputted, soundexchange.com would say that 'no results were found'. A proxy was also used in this algorithm to disable image loading, search in incognito mode and use anti-detection features to optimize searching. From here, using the ISRC code, the artist name, song title and recording ID were extracted, returned as a list and saved to a new CSV file.



scraped_data1

None	None	GBKQU1524393	TRUE
None	None	ITY701800108	TRUE
None	None	US83Z1106885	TRUE
Olivier Py / Birds of Paradise	Punk Prototype, Pt. 2	FR9W11700196	TRUE
Newban	Find a Place to Live	GBEQT1203283	TRUE
Father	Ghosts	BEY920901807	TRUE
Charlotte Someone	Another Fine Day	USLZJ1956668	TRUE
None	None	GBDDN1400602	TRUE
None	None	NLCK41065531	TRUE
Golden Grand	Too Club For Saks	QZPLS2147312	TRUE
Demolish Beatz	I Use What's Left	USXQS1923227	TRUE
Spanless	The Essence of Truth	US83Z1448772	TRUE

Content-Based Recommender

The content-based recommender leverages six audio features – danceability, energy, valence, tempo, acousticness, and liveness – to identify and recommend songs that share similar characteristics with a given input song. Using the sigmoid kernel, a nonlinear similarity measure, the system calculates pairwise similarity scores between the feature vectors of the input track and all other tracks in the merged dataset. The process begins by standardizing the audio features using StandardScaler to ensure that all features contribute equally to the similarity calculation. Once a song title is provided as input, the system identifies its corresponding index in the dataset and computes the similarity scores. The top matches are then ranked by their similarity score, and the top recommendations are returned as shown in the table below. The results table provide details such as song title, artist, ISRC, key audio features, and similarity score. This recommender is especially effective for users seeking music discovery based solely on inherent song characteristics as the dataset does not include user data.

Recommendations based on the song: Dream of a Machine

	Title	Artist	ISRC	Danceability	Energy	Valence	Tempo	Acousticness	Liveness	Similarity Score
0	I Got It Made (Re–Recorded / Remastered)	Various Artists	USA370956824	0.666	0.751	0.85	190	0.0156	0.0999	0.907972
1	Superstars	Styles Of Beyond	US3260400034	0.505	0.865	0.14	216	0.00259	0.0563	0.906435
2	Midas	Baselinez	USA2P1680614	0.72	0.827	0.563	194	0.00468	0.248	0.905506
3	Computer Glitch	Positive Postulate Records	QZMHM2093554	0.702	0.709	0.639	190	0.0109	0.0589	0.902388
4	Dice Roll	Spiral Helix	QZNWY2352808	0.563	0.782	0.812	192	0.117	0.162	0.897396
5	BANGBAP	Bob Catt The Legend	QZ5FN2082467	0.707	0.742	0.963	180	0.116	0.0472	0.896668
6	Reggaeton Backing Track – A minor	Gene2020	QM6MZ2061016	0.686	0.735	0.735	180	0.00283	0.333	0.89225
7	Te Cogio	David El Embajador	ITJ871800178	0.724	0.893	0.94	176	0.171	0.154	0.889094
8	Super Natural	Coe	UKHYY2200017	0.662	0.416	0.0397	200	0.0284	0.139	0.883567
9	Something from the Old School	DJ Tempo	GBKPL1520887	0.707	0.706	0.599	174	7.25e–05	0.0912	0.881718

This output table is based off the input song, "Dream of a Machine" by Zagar. This song is a smooth and atmospheric electronic track that encompasses steady beats and layered melodies, creating a dreamy yet mechanical feel. After inputting the song title into the recommender function that top results are displayed, with "I Got it Made (Re-Recorded/Remastered)" by Various Artists as the most similar (91%). This recommendation makes sense since these two songs have similar values for danceability (0.59 vs 0.66) and energy (0.522 vs 0.751).

Matrix Factorization Recommender

To improve our content based recommender system we opted to also include matrix factorization. The matrix factorization recommender uses implicit feedback, primarily user play counts for these results, to generate personalized song recommendations. Utilized an Alternating Least Squares model that decomposes the user-item interaction matrix into latent factors. These factors represent the user preferences and the song characteristics. The recommender can then identify user behavior patterns and predict songs based on their listening history. The interaction matrix is then normalized using BM25 weighting, which adjusts the play count data by reducing the dominance of popular overplayed songs, which could skew the recommendations. This helps the recommender not be too oversaturated by overplayed music. By normalizing the data with this method, the BM25 weighting helps balance the influence of different users and songs, ensuring more independent artists are given fair consideration. Once a user ID is provided, the ALS model computes scores for all songs by comparing the latent user preferences with the song features. Then the top songs are then uploaded and displayed in the output table.

	Title	Score
0	Remember	0.993934
1	Logman's Beak	0.992078
2	Wrong Side Up	0.991583
3	Up in Smoke	0.991453
4	Let You Go (feat. Yves Paquet)	0.990825
5	The Day I Lost Everything	0.990816
6	Wisdom of the Universe	0.990644
7	Stay in Love	0.990468
8	Bormaz - (maxi Version)	0.989612
9	Bring your Love	0.989374

This output table combines the content based scores with the matrix factorization scores creating a more accurate recommender system as shown above. With this combined approach the top similarity score is 99%. Again, looking back to the recommended song "Dream of a Machine" by Zagar the top recommended song is "Remember" by Matador. Both tracks are rooted in electronic music, featuring synthesizers and steady rhythms. They share an emphasis on layered soundscapes and a futuristic aesthetic which makes sense that it is the most recommended song.

Conclusion

In conclusion, we attempted to build a music recommender that would recommend the top 10 similar songs on Spotify to the user. Our original audio features dataset on Kaggle did not have the artist name or song title so we used web scraping via soundexchange.com to create a csv file containing the relevant information. Using a combined score from matrix factorization and content based recommendation, we then recommended the top 10 similar songs to the user. Our best performing result had a top similarity score of 0.992479 (on a scale from 0 to 1). In the future, if we were to scale our project, we could develop a user-friendly interface for users to input song preferences, incorporate additional data sources to scrape such as Apple Music and optimize our recommendation algorithm even further to reduce computation time.

The code for our project is submitted separately in a Jupyter Notebook file on Canvas and can be located at this link: https://github.com/kirajackson/134-Final-Project/blob/main/134%20Project%20Code%20Recommender.ipynb.