VISUALIZING THEMES ACROSS THE TOP 40 MUSIC CHARTS THROUGH TIME USING NATURAL LANGUAGE PROCESSING ON SONG LYRICS

Nicholas Hatcher Kevin Holmes Eric Wissner Udit Dasgupta

CSE 6242 Data & Visual Analytics - Fall '22 - Team 12

MOTIVATION

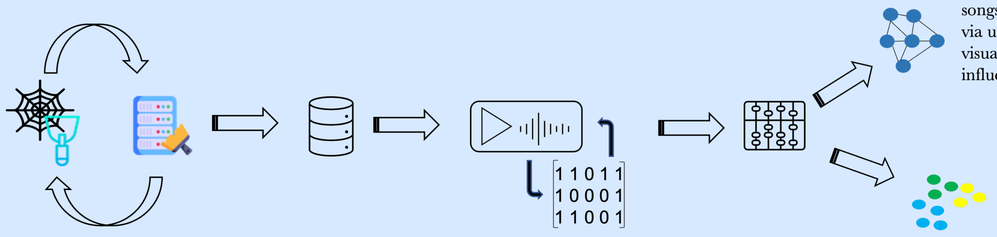
As music lovers we wanted to do a deep dive into the into the most popular music in the United States of the last 70 years and understand the most prevalent themes and the relationships between artists. Our project hopes to give ourselves and other music lovers a rich, visual understanding of popular music.

Visualization Approach

The goal of this project is to understand the most prevalent topics, themes, or sentiments expressed in top 40 music across time. We combed Natural Language Processing, machine learning and visualization techniques to create an intuitive dashboard to analyze lyrics, audio characteristics, and relationships between artists.

The primary innovations of our approach are:

1. A temporal visualization of lyrics sentiment.
2. Leverage & combine unsupervised learning techniques to provide a visualization framework that is domain agnostic.



Adjacency matrix, representing pair-wise relationships between songs. Visualization of matrix via undirected graph to visualize musical themes and influences over time.

US Billboard Top40 music rating from 1955-2019 obtained using web scraping and data cleansing performed using regular expressions, string methods, and off-the-shelf data wrangling tools.

Ranking, artist, raw lyrics, track name, and audio features for each track stored in SQLite database. ~130,000 records with each record representing an appearance on the Top40 listing.

Tokenization of lyrics and construction of Term-Document (TD) matrix represents the frequencies of each word in each song, thereby creating a mathematical representation of the corpus.

Transform TD matrix into a Term Frequency-Inverse Document Frequency matrix (TF-IDF). Stop words were not removed as high frequency words related to harmonization, or other vocalizations are properly accounted for in standard libraries.

Latent Dirichlet Allocation (LDA) to capture "hidden" themes in artist music over time

Data

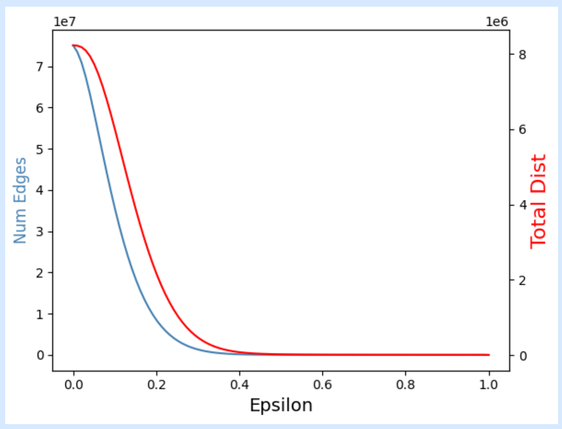
Data Source	Billboard Top40 (https://top40weekly.com)	Spotify
Obtaining the Data	Web scraping using Python BeautifulSoup	API calls
Cleansing the Data	Using regular expressions, Python string methods, and other off-the shelf data wrangling tools such as OpenRefine to match Artist/song from Billboard data to respective lyrics from Spotify	
Characteristics	~180,000 records representing track listings and rank in top 40 across time	~18,000 unique songs along with lyrics and 12 Spotify "sonic" features
Storage	Integrated into SQLite8 database (~80 mb on disk)	

Experiments & Results

TF-IDF Model Parameter Tuning

TF-IDF was used as an intermediary step towards creating a useful visualization. We experimented with measuring song similarity using Euclidean distance and cosine similarity and found that use of cosine similarity yielded better results in downstream visualization.

As initial graph consisted of ~7MM pair-wise relationships, the size of the graph was managed via definition of a minimum threshold, epsilon, to remove weakly connected edges and enforce sparsity. Empirically we found significant reduction in edges at epsilon = .3; reducing the number of edges to ~1.3MM. While still large we felt this allowed for relative sparsity while also providing a robust dataset for visualization.

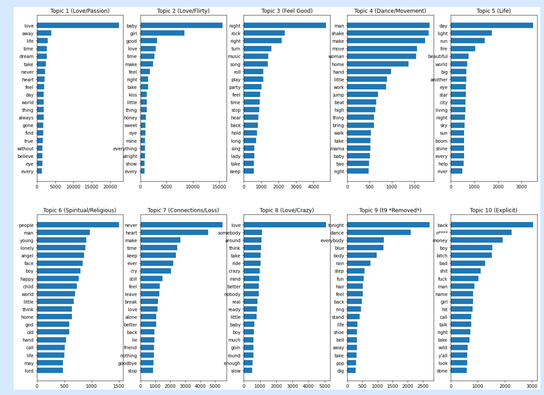


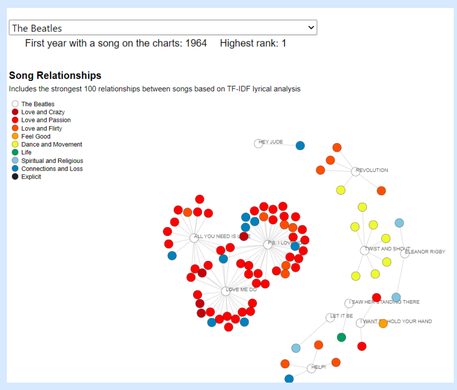
Experiments & Results

LDA Model Parameter Tuning:

The main experimental tuning of LDA was labelling the latent topics that the algorithm was discovering. LDA uses a bag of words (BOW) count vectorization as an input to the algorithm.

Labelling topics involved manually examining a combination of the most representative words for each topic (an output from the LDA algorithm) and looking at the top 20-30 songs that had the highest distributions on a respective topic. We decided to use 9 of the 10 topic outputs from the LDA (the 10th appearing to have no coherent meaning). The final topics were labeled as Love/Passion, Love/Flirty, Love/Crazy, Feel Good, Life, Dance/Movement, Spiritual/Religious, Connections/Loss, Explicit. We stopped iterating once a coherent trend emerged in the songs and timeline data.





Comparing TF-IDF and LDA:

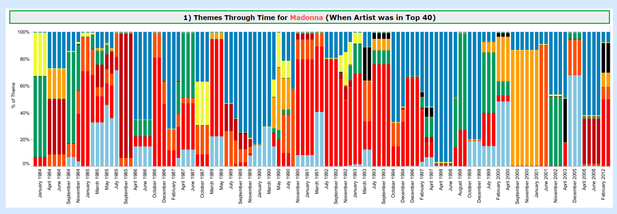
A convenient aspect of our visualization is that one can visually verify if the two algorithms give similar results.

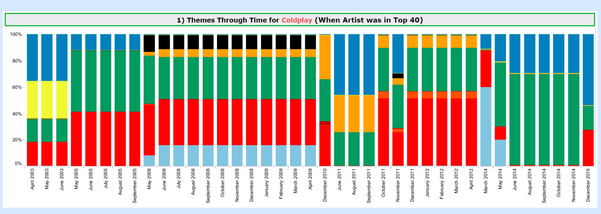
As shown on the left, the songs with high similarity scores also have similar topic classifications as per LDA.

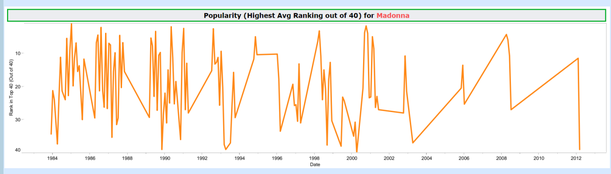
Through extensive inspection of multiple artists, we find that the two algorithms supplement each other's results providing a robust ensemble technique.

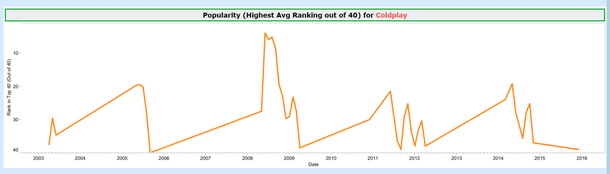
Madonna vs Coldplay: Commercial success trends with diverse and evolving themes

On comparing a consistently successful long term artist such as Madonna with moderately successful artist like Coldplay, we can hypothesize that artists with more diverse and evolving themes tend to do better commercially over the long run. This is consistent with the findings of Berg (2022) which noted that novelty of styles helps sustain success but makes it harder to break through in the first place.









Madonna vs. Beyonce: Comparing Major Pop Star Icons of different generations

An interesting trend we found was that of comparing two pop icons from different eras. Madonna and Beyonce both had consistently dynamic theme and popularity charts, but Beyonce had more interconnected songs in the graph. A possible research question could be if originality in lyrics has reduced over time leading to higher similarity in lyrics