Music Analysis & Recommendation Engine (M.A.R.S.)

Simon Fraser University, Big Data Programming II

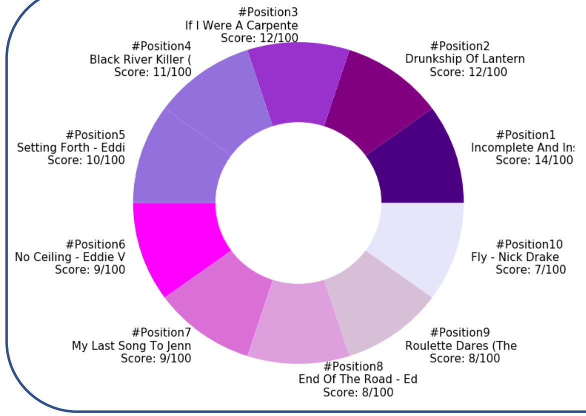
Kashish Kohli
Kanksha Masrani

Problem & Approach

- In 2018, the global music industry was worth USD 130 Bn + and is estimated to grow faster than ever fueled by the rise of paid streaming services.
- Numerous advantages vs challenges.

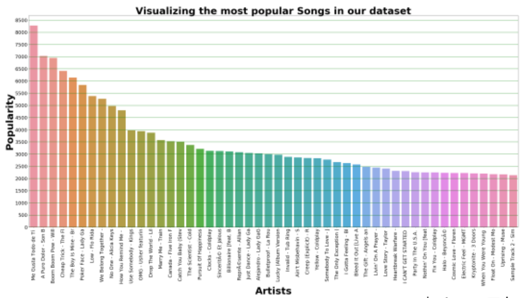
- We developed a Popularity Based Recommender and a User Similarity based Collaborative Filtering Model.
- Song popularity prediction using 6+ ML algorithms.
- Music Sentiment Analysis by region.

Data Collection



Recommendation Engine

- Million Songs Dataset with 2 million songs and 75,000 user profiles.



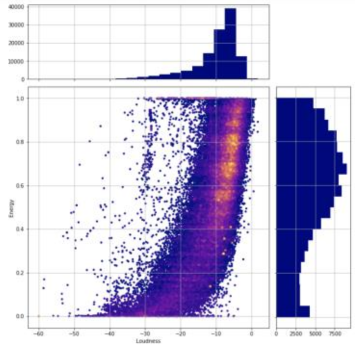
- Calculated Recommendations for the User Similarity model by using Jaccard Index to plot the Co-occurrence matrix.

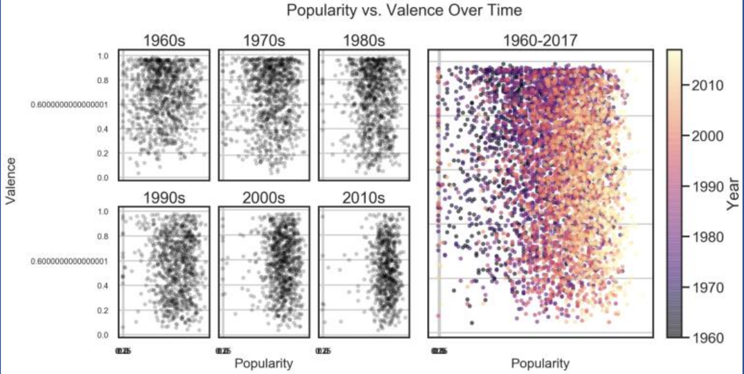
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Included dual functionality of suggesting songs based on history of the user as well as based on a single song alone.

Song Popularity Prediction

- Amalgamation of Spotify and MSD Dataset with 120,000 records of music features and metadata.
- Analyzed trends to map success of music in the present and future

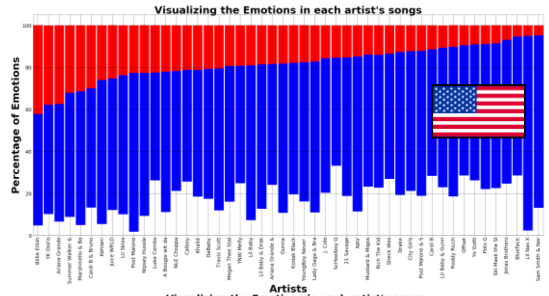


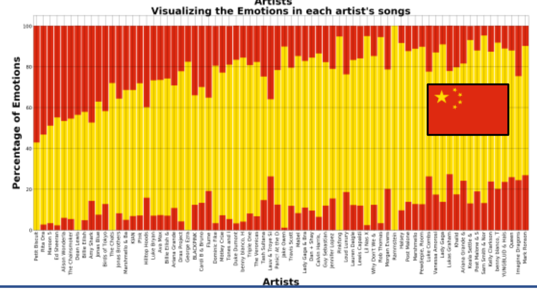


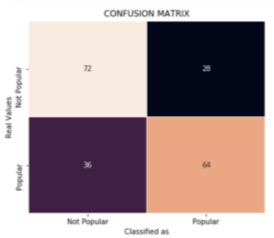
- Machine Learning Models using Gridsearch for Model Tuning:
  - KNN Clustering : 62%
  - Support Vector Classifier : 65%
  - Adaptive Boosting : 65.5%
  - Logistic Regression : 66%
  - Convolutional Neural Network : 67.5%
  - Random Forest : 70%

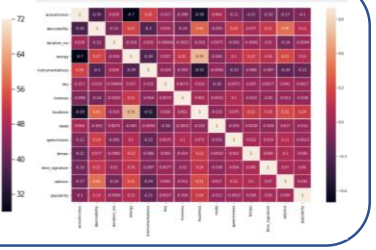
Global Sentiment Analysis

- Extracted data from iTunes RSS Feed Generator to get top 100 hit songs for multiple countries.
- Calculated the Lexical Richness of each country's top songs and performed sentiment analysis using NLTK Sentiment Vader, GoogleTrans and GoSlate.

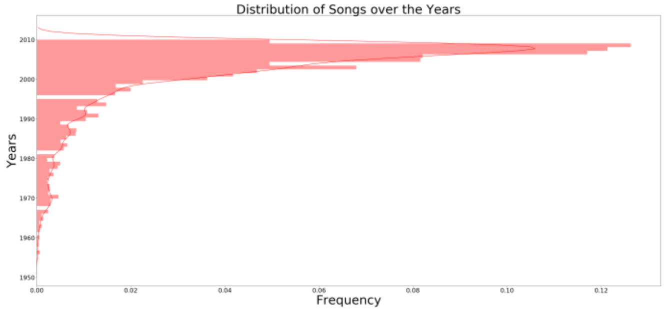








Recommendation System



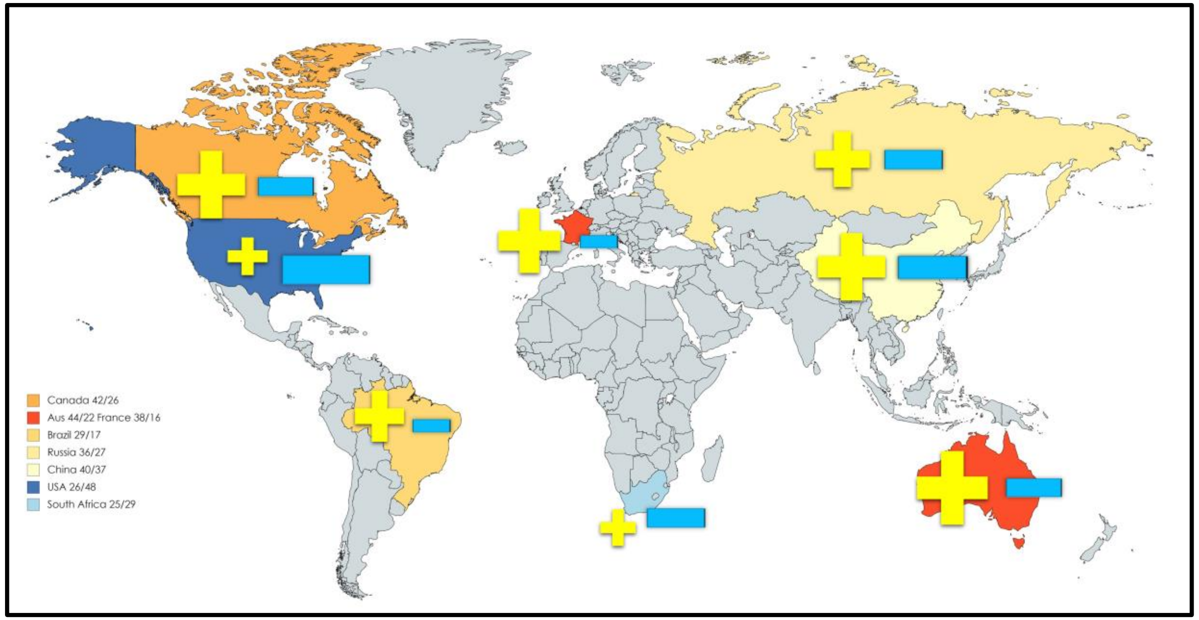
Popularity Based Recommender

	Song Name	Similarity Score	Position
2309	Me Gusta Todo de Tí - Luis Alfonso Lizárraga	34	1.0
66	A Puro Dolor - Son By Four	26	2.0
657	Cheap Trick - The Flame	23	3.0
3975	Use Somebody - Kings Of Leon	23	4.0
2818	Poker Face - Lady Gaga	22	5.0
3576	The Boy Is Mine - Brandy Norwood	21	6.0
3702	The Scientist - Coldplay	21	7.0
476	Boom Boom Pow - Will.I.Am Fergie	19	8.0
1605	How You Remind Me - Nickleback	19	9.0
2218	Low - Flo Rida	18	10.0

User Similarity based Collaborative Filtering Model

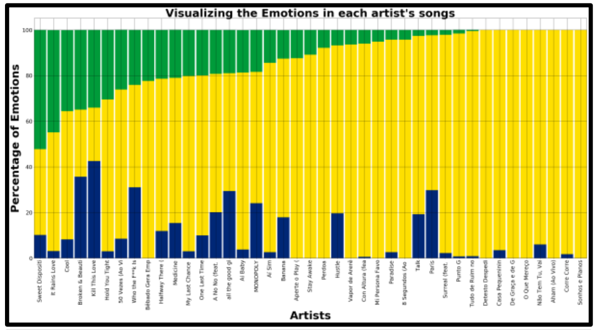
	Song Suggestions	Similarity Score	Position
0	Isolation - Joy Division	0.030351	1
1	Transmission - Joy Division	0.029383	2
2	Shadowplay - Joy Division	0.029157	3
3	Digital - Joy Division	0.027941	4
4	The Stranger Song - Leonard Cohen	0.021234	5
5	Dead Souls [Re-mastered] - Joy Division	0.020127	6
6	The Killing Moon - Echo And The Bunnymen	0.016971	7
7	Damaged Goods - Gang Of Four	0.016238	8
8	Friction (LP Version) - Television	0.014653	9
9	This Charming Man - The Smiths	0.014265	10

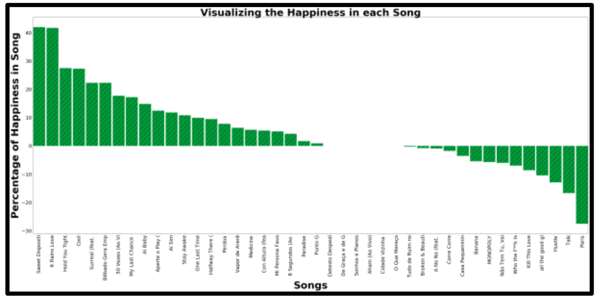
Global Music Sentiment Analysis



Workflow

1. Fetch JSON format data from iTunes RSS Feed Generator
2. Fetch lyrics of these songs from Genius.com using Pypi wrappers
3. Analyze unique wordcount to calculate lexical richness.
4. Translate the song (if required) using GoogleTrans/GoSlate/Pypi wrappers
5. Perform lyrical sentiment analysis using NLTK Sentiment Vader and calculate the percent of positive/neutral/negative lyrics to ascertain song trends in each country.
6. Plot the positive/negative percentage of songs heard in the country.





Data Collection + Cleaning

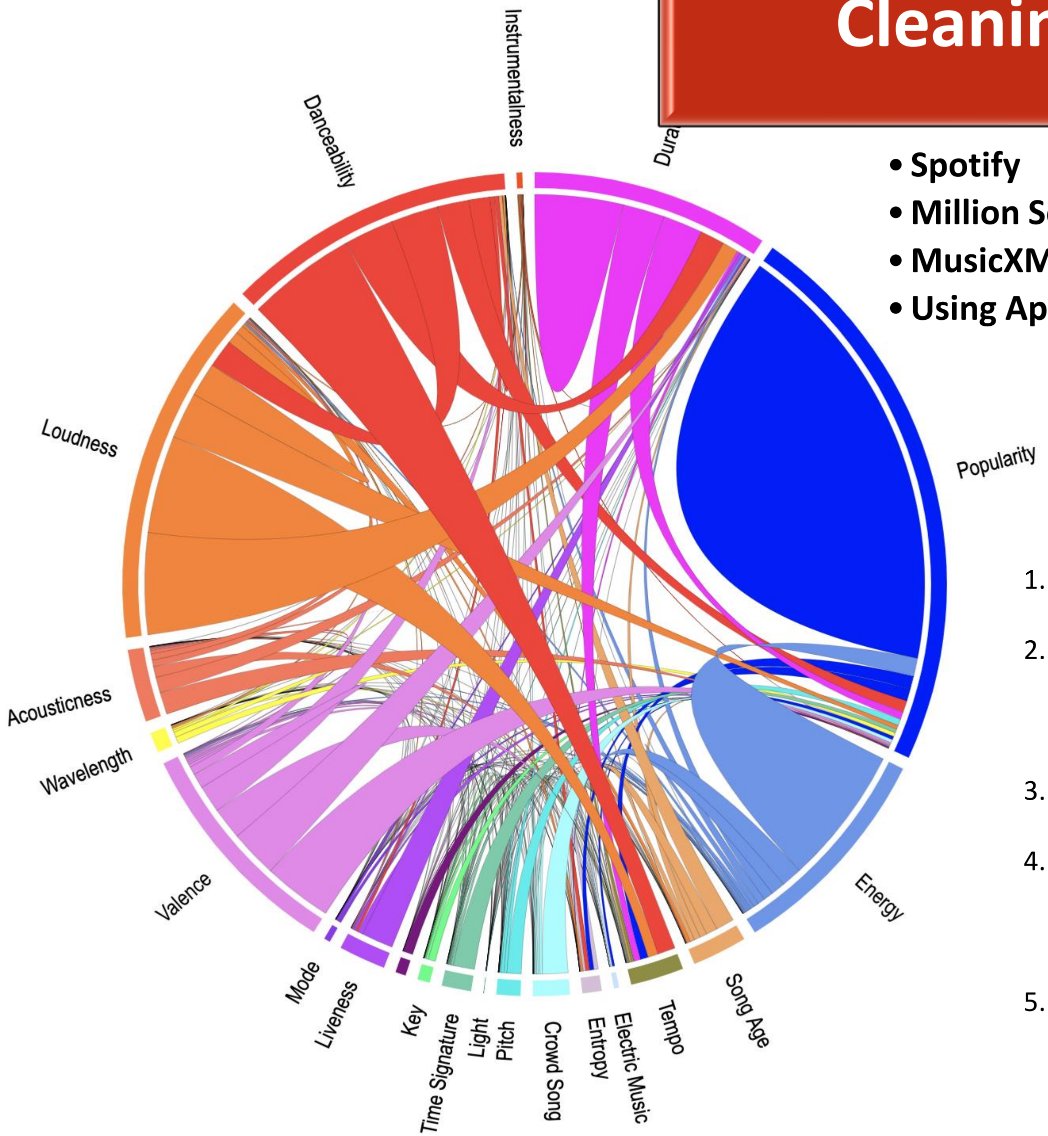
- Spotify
- Million Songs Dataset
- MusicXMatch
- Using Apache Pig

Exploratory Data Analysis

- Detecting Outliers
- Plotting Correlations

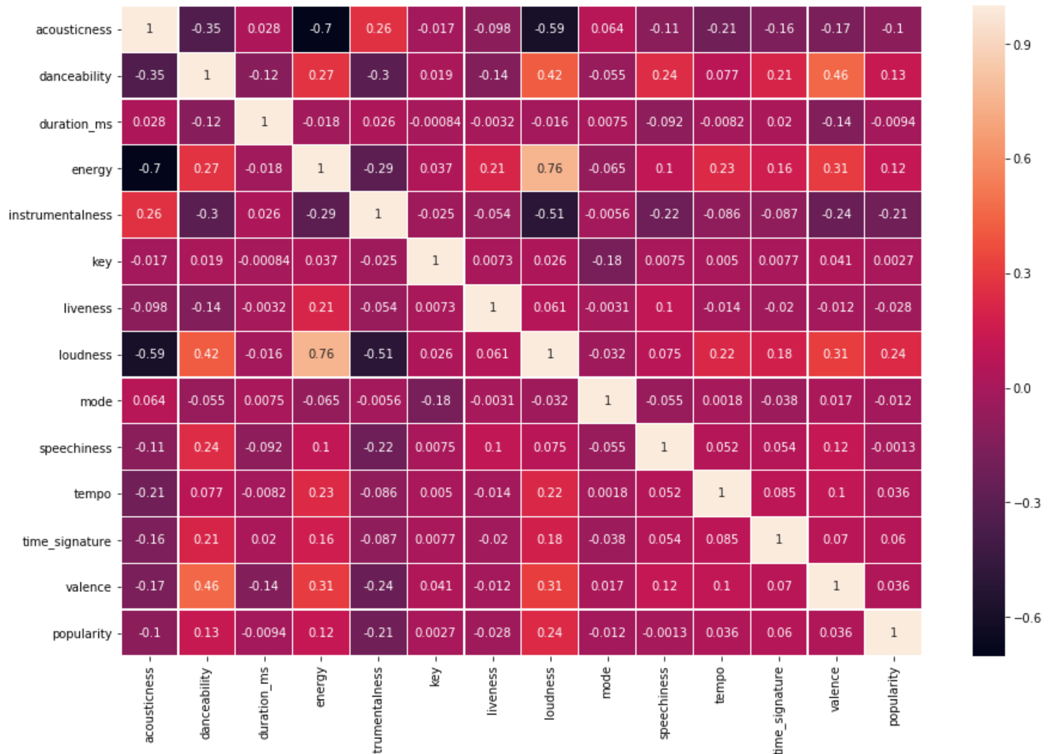
Data Manipulation

- One Hot Encoding
- Categorization into popularity classes



FINDINGS

1. Popularity is highly related to Loudness, Danceability & Energy
2. Popularity drops sharply with increase in Instrumentalness, Liveness, Mode and Speechiness.
3. Valence also is highly related to Loudness, Danceability & Energy
4. Valence drops sharply with increase in Instrumentalness, Duration, Acousticness and Liveness.
5. Energy and Loudness with Acousticness and Instrumentalness present the worst combinations



Split Data

With & without Outliers

- **Train test Split**
- **80% Training Data vs 20% Test Data**
- **64000 vs 16000**

Design & Test Models

Test Results

- **Confusion Matrix**
- **Classification Report**
- **ROC Curve**

KNN Clustering
- Accuracy 0.63
- With Cross Validation 0.62

Adaptive Boosting
- Accuracy 0.65

Support Vector Classifier
- Used Gridsearch to get optimum parameters
- Accuracy 0.65

Convolutional Neural Networks
- Used Gridsearch
- Accuracy 0.67

Random Forest
- Used Gridsearch
- Accuracy 0.7