Project Type: Analysis

Group: Crystal Contreras, Tavis Sotirin-Miller

Overview

Analyze yearly song releases in ‘Top [Genre]’ charts recorded by Billboard (American entertainment media company) for various genres, using song data from Spotify’s database. A songs’ ranking in the chart will determine ‘popularity’ for that year, and a categorical feature of popular or not (based on being in a top chart or not for that year) will be used as target data.

Yearly top songs will be scraped from Billboard’s website, along with listings of other songs released during the same year, which will be compiled into a working data set.

Analysis of the Spotify song features will be performed, and the dataset will be optimized for prediction/recommender systems.

Various classifiers and predictive models will be tested on an ~80/20 split of the dataset to predict a years’ generalized song features, as well as to attempt to classify testing songs as popular. Results will be discussed, and the best method/model will be presented.

Analysis Approach

Task 1: Pre-Processing

Subtask A: Collect Data

- Scrape Billboard’s historical top genre charts off website for popular song list
- Create ‘unpopular’ songs list per year per genre
- Retrieve Spotify song data for collected popular and unpopular songs
- Retrieve release date for each song from Spotify using corresponding album information

Subtask B: Data Pre-Processing

- Convert categorical features into dummy variables
- Flag tracks from top charts that were not released in the same year for potential removal during analysis
- Convert release date into ‘distance’ from January 1 of corresponding chart year, or a similar numeric metric
- Clean up encoding / address any missing data values (plan to remove corresponding records)
- Combine song lists with original rankings from top charts to get a ‘popularity’ feature
    - Consider creating buckets for popular instead of using direct ranking

Task 2: Exploratory Phase

Subtask A: Explore song data set

- Analyze data for outliers
- Analyze data for correlated features
- Analyze merits in features against the selected features the model outputs (e.g., duration would most likely get removed)
- Use PCA or Forward/Backward Selection to determine which features are most important to determining popularity of the song per genre

Subtask B: Cluster analysis

- Use kNN to cluster buckets of top 100 tracks (e.g., popularity clustering)

Task 3: Classification and Prediction

Use kMeans for clustering;
kNN is classification

Subtask A: Classification

- Build various models (Decision Tree, Naïve Bayes, and potentially others) using popular and unpopular song data to classify an input song (target variable will be popular vs. unpopular)
    - Build off historical data prior to 2019, using 2019/2020 as testing data
    - Separate build off training/testing data split by year, with classifier only looking at a single year a time
        - Compare results to see if song trends have changed enough over the years to skew the model when using entire history as training data
- Analyze and compare results between models and discuss historic song trends impact on the models and highest accuracy model

Subtask B: Prediction

- Build predictive models (Linear Regression, and potentially others) using popular song data to predict the next years generalized song vector
    - Build off historical data of only top chart (popular) songs prior to 2020
        - Compare results with actual averages and ranges of popular song vectors in 2020
- Analyze and compare results between models and identify highest accuracy model

Data Schema and Size

Expected dataset size will average 70 popular and unpopular songs per year per genre, with an average of 10 years of history. Approximately 1400 songs per genre, for an expected 5 genres. Approximately 7000 songs total.

Anticipated Feature Set

Categorical:

- Genre
  - Country
  - Jazz
  - Latin
  - Pop
  - R&B
- Popular
  - On top chart lists
  - Not on top chart lists
- Key
  - A – G# (numeric from 0 – 11)
- Explicit
  - Contains explicit content
  - Does not contain explicit content
- Mode
  - Major key
  - Minor key

Numeric:

- Chart ranking
- Acousticness
- Danceability
- Energy
- Duration
- Instrumentalness
- Valence
- Tempo
- Liveness
- Loudness
- Speechiness

Plan for Evaluation

Classifiers will be tested on both popular and unpopular song vectors and results and accuracy will be discussed among the various models. Predictor results will be compared to actual data (against mean vectors and normalized vectors for the years tested). Again, results and accuracy will be discussed between the various models.

Model results will be compared via statistics of performance, including accuracy, error rate, f-1 score, etc.

Work Distribution

Task 1: Tavis

Task 2: Crystal, Tavis (Done independently and results discussed and combined)

Task 3a: Crystal

Task 3b: Tavis

End results will be reviewed and discussed by both group members for final analysis