Predicting Release Genres from Lyrics using Music Releases Webscraped from Bandcamp

Sam Roberts-Baca & Thomas Tellner
University of Denver, COMP4448 Fall 2021

EXECUTIVE SUMMARY

A new, unsigned music artist desires to uncover what similarities exist between music releases from various genres in terms of their lyrical content. Specifically, he would like to know if the genre of a piece of music can be predicted by its lyrical content. Accordingly, these results would guide him regarding what lyrical themes and topics are prevalent in various kinds of music. Through using Multinomial Naive Bayes and KNN Classifier models, we attempted to see how well we could predict release genres by lyrics for a set of webscraped musical releases from Bandcamp, an online independent music distribution platform. This project is of significance primarily for unsigned independent artists who are interested in self-releasing their music online and music researchers who want to learn more about larger cultural themes that emerge through linguistic analysis of contemporary, independently released music.

DATASET AND MOTIVATION

This dataset was collected by web scraping the Discover page on Bandcamp.com, an online independent music distribution platform. The Discover page and its various filters are shown in Figure 1.

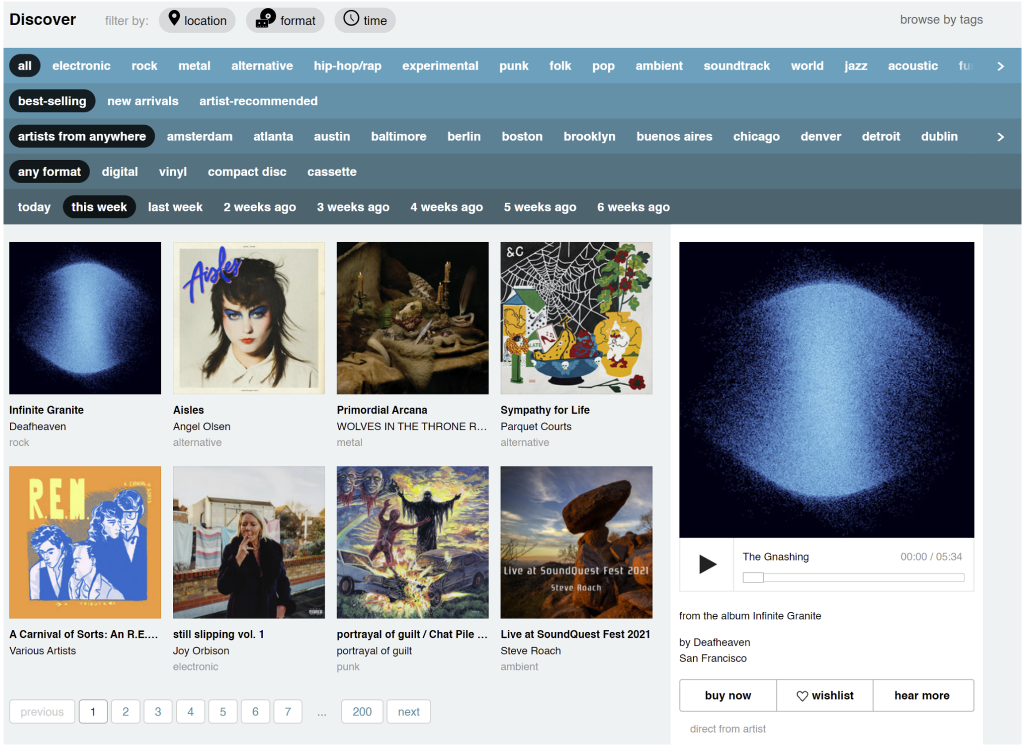

Figure 1. Bandcamp's Discovery Page

We wanted to analyze the lyrical content of various albums released over the last six weeks on the platform in various genres and locations of independently released music, to determine how various genres and locations of music released in the English language compare in terms of their lexicon used.

The metadata of the data collected is as follows:

Release Genre - The search genre parameter. For example, rock, pop, etc.
Release Sub-Genre - The search sub-genre parameter. For example, for the rock genre: indie rock, psychedelic rock, etc.
Search Format - The search format parameter. Valid inputs include:

- All Formats
- Digital
- Vinyl
- CD
- Cassette

Search Week - The search week parameter. Valid inputs include:

- Today
- This Week
- 2 Weeks Ago
- 3 Weeks Ago
- 4 Weeks Ago
- 5 Weeks Ago
- 6 Weeks Ago

Search Category - The search category parameter. Valid inputs include:

- Top - The best-selling releases for a given search
- New - The newest releases released for a given search
- Rec - The most recommended release by artists for a given search

Release URL - The URL corresponding to the release on Bandcamp.
Scrape Date - The date of the web scrape of the release.
Release Title
Artist Name
Artist Location
Release Date
Tags - A list of keywords associated with the release.
Track Info - A list of tracks containing the following information for each track:

- Track Name
- Track Number
- Track Duration
- Track Lyrics

All Lyrics - A string containing all lyrics for a given release.
Number of Tracks - The number of tracks (songs) for a given release.

On August 20th we used our web scraping tool to scrape 5165 releases on Bandcamp. The genres we searched for were rock, alternative, hip hop/rap, experimental, pop, acoustic, country, blues, jazz, R&B / soul, and reggae. The *best selling* and the *newest* albums for each genre were collected for 1) this week, 2) three weeks ago, and 3) six weeks ago. All formats were considered. For each query, the first thirteen pages of results were scraped, leading to approximately 104 releases scraped for each query (filtered by genre, week parameter, and top/new category).

Python was used to create the Bandcamp web scraper and to analyze the resulting data.

The input variable we used for our analysis was the “All Lyrics” column. The output variable we used for our analysis was the “Release Genre” column.

RESEARCH QUESTION

An independent artist would like to know what kinds of music in terms of its lyrical content is being released in that artist's target genre and location. The artist would like to know what is trending lyrically so that he/she can know what commonalities independently released contemporary music share.

The inputs of our analysis are lyrics, and the output of our analysis is a release genre.

Our research questions are as follows:

1. How well does a Naive Bayes algorithm predict a music release genre from its lyrical content?
2. How well does a KNN-Classifier Algorithm predict music release genre from its lyrical content?

DATA PREPROCESSING

The following process was used to preprocess the webscraped Bandcamp dataset. First, the "Release Genre" and "All Lyrics" columns were selected from the larger data set. All albums with "Release Genre" marked as all "All" (where "All" was the search parameter for the web scrape) were then removed from the dataset. All release rows where lyrics were not in English were manually removed. The "All Lyrics" column was cleaned in the following manner:

- Any tags, hashtags (#), and mentions (@) were removed
- URLs were removed
- Tags were removed
- English stop words were removed
- Punctuation at the end of each word was removed
- All words were made lowercase

After preprocessing 986 releases were obtained for analysis. Figure 2 shows the breakdown of preprocessed releases by genre. Figure 3 shows the average number of words present in the lyrics for a release by genre. Interestingly, releases classified as hip hop/rap tended to have the most words present in the lyrical data, while jazz releases had the lowest amount of words present.

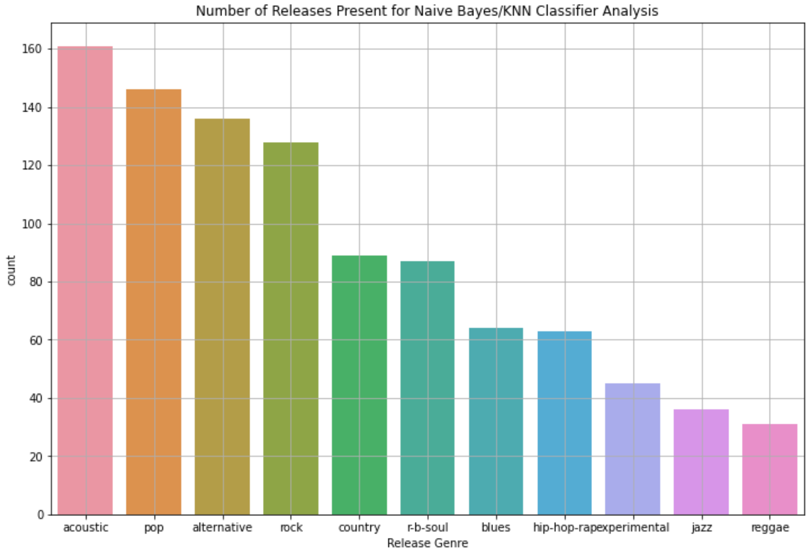

Figure 2. Breakdown of Preprocessed Release Genres

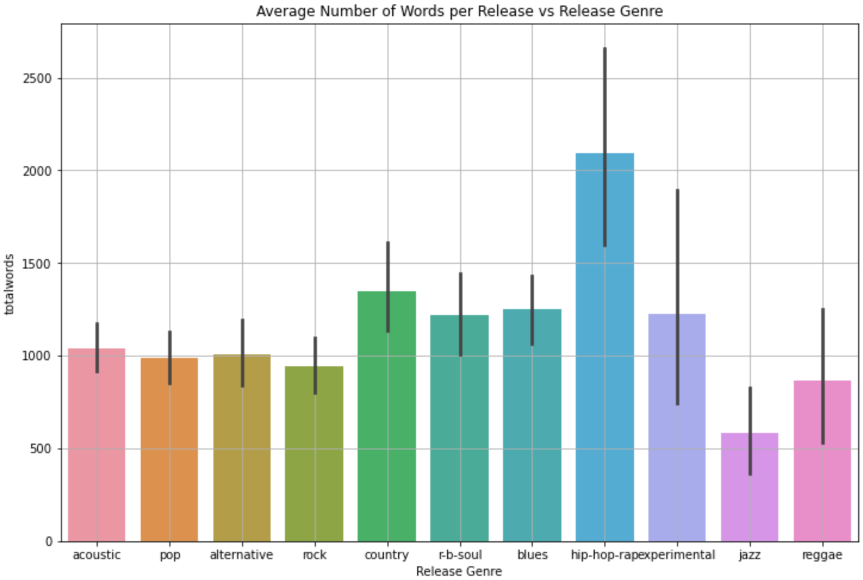

Figure 3. Average Number of Words per Release for each Release Genre

MODEL IMPLEMENTATION

The two specific models we picked to test against one another in this project were picked by way of a brief review of NLP-related content for the most-often mentioned models used similarly. Two of the most popular models were Naive Bayesian and KNN and these were the models we chose to implement and compare as part of our research questions.

We also opted to use the models “pre-built” in the sklearn package rather than implementing two models from scratch and then comparing them. By using the sklearn models we avoided any issues with any particular coding errors that might occur in an unvetted model and therefore give the other model an undue advantage. The sklearn models have been widely used and vetted by the user community and as a result using them “levels the playing field” between them when they are both applied to the same dataset.

In both cases, we implemented a train-test split using a test size of 40%. We also narrowed the genres down to 2: Country and Rap.

1. Naive Bayes: We implemented the Multinomial Bayes algorithm from sklearn
2. KNN: For KNN we implemented the sklearn algorithm with defaults (i.e. k=5)

The initial results are displayed in the table below. After the first run on of the models we applied GridSearchSV for tuning the hyperparameters. The hyperparameters that were tuned were as follows:

1. Naive Bayes:
    a. Min_df: this sets the frequency “floor”. Words occurring less often than this number will be disregarded.

    b. Ngram_range: this sets the “blocks” of words to analyze, i.e. a value of 1 means 1 word strings, whereas a value of three would be three words. Examples: “love” versus “I love you”

2. KNN:
    a. N_neighbors: this is the core hyperparameter for this algorithm, defining how many nearby points to check and calculate the majority label values.
    b. Min_df: see above
    c. Ngram_range: see above

The results of fitting the two models were as follows:

Table 1. Results for Two-Genre Multinomial Naive Bayes and KNN Classifier Models

	Training Acc. (Before tuning)	Testing Acc. (Before tuning)	CV Score	Training Acc. (After tuning)	Testing Acc. After tuning
Naive Bayes	.95604395	.81967213	.814035088	.9010989	.78688524
KNN Classifier	.84615385	.75409836	.714035088	.8681318	.70491803

DISCUSSION & NEXT STEPS

In terms of the effectiveness of using Naive Bayes to predict genres from lyrics, we concluded that Multinomial Naive Bayes is not very accurate for predicting release genre given lyrics where there are many output genres to choose from (11 genres), given our resulting cross-validated accuracy of .21. Multinomial Naive Bayes is far better at predicting release genre given lyrics where there are only two output genres to choose from, as our resulting cross-validated accuracy for this model was .81. It should be noted that the model used to predict two genres used “Hip Hop/Rap” and “Country” as the genres selected, and the final accuracy for the model when using other combinations genres varied, but was still generally better at predicting genres than the complete 11-genre model.

On the whole, our KNN Classifier performed similarly to our Naive Bayes for predicting “Hip Hop/Rap” versus “Country” lyrics. With a resulting cross-validated accuracy of .71, this model was slightly worse at predicting release genre given lyrics where there are only two output genres to choose from.

For the two genre model, both our Naive Bayes and KNN Classifier models suffered from overfitting as the training set prediction accuracies were somewhat high compared to the testing set prediction accuracies, roughly 15% higher for the Naive Bayes model and 10% higher for the KNN Classifier model.

In the future, we would like to look at how accurate the NB and KNN models are for comparing any two, three, four genres, etc., by comparing the averages of NB/KNN accuracy for each combination of any number of genres.

Given the influence songs and music have in our lives, the development of this kind of textual analysis and the development technologies leveraged in the future are important topics for further study.