NLP-Based Genre Classification of Music

Benjamin Prud’homme

Abstract

In this project, I explore the association between song lyrics and musical genre through natural language processing. I develop a model for categorizing songs by genre based on their lyrics, which uses a bag-of-words model with TF-IDF vectorization and a logistic regression classifier. When trained on a set of songs from seven different genres, this model performs at a reasonable level of accuracy, and succeeds in identifying several features that are often associated with specific genres.

Introduction

The concept of musical genres is quite loosely defined. They are generally thought of as sets of stylistic conventions, cultural influences, and other artistic elements that commonly and reliably distinguish songs from each other. Lyrics play a large part in this categorization: the words that songwriters use and how often these words are repeated are often seen as quite characteristic or even stereotypical of certain genres. Today, there is an ever-growing library of music available on the internet, which prompts the development of algorithms to allow users to browse these databases for music they like. With this in mind, my goal in this project is to develop a classifier that can interpret songs as belonging to certain genres based on the lyrics alone.

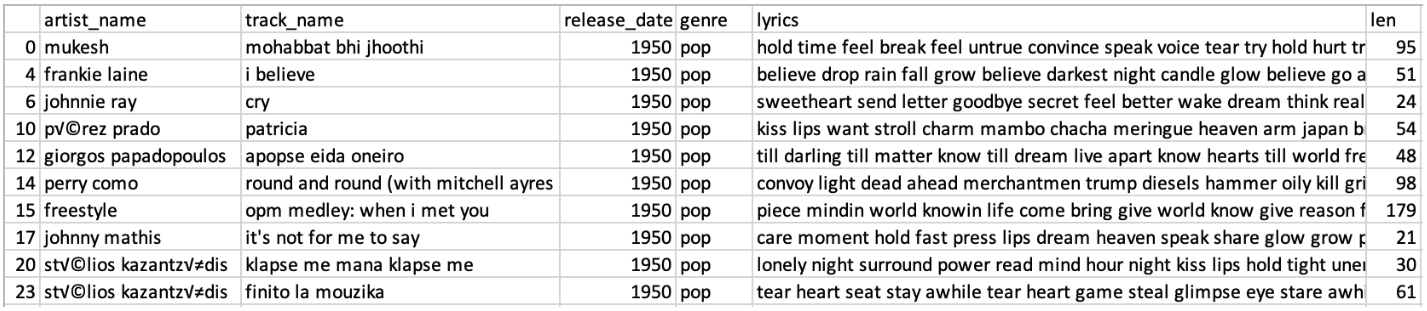
Existing Work

Various methods have been developed to classify music based on lyrical content in recent years. In his article “How We Used NLTK and NLP to Predict a Song’s Genre From Its Lyrics,” Dilyan Kovachev details his process of experimenting with multiple machine learning algorithms to classify music by genre. His best model achieves an accuracy of 50%. Michael Fell and Caroline Sporleder make use of n-gram models along with identification of attributes such as vocabulary, structure, and style to perform detection of genre. Luan Moura, Carlos Forster, Emanuel Fontelles, Vinicius Sampaio, and Mardonio Franca go a step further by implementing a topic model to visualize trends in lyrics across genres over time.

Data

The dataset I used for this project was made available by de Moura, Forster, Fontelles, Sampaio, and Franca (studied in their paper “Temporal Analysis and Visualization of Music”). It contains lyric data for 28,372 songs across 7 genres (blues, country, hip hop, jazz, pop, reggae, and rock) that were released between 1950 and 2019. They obtained this data through the *spotipy* (for querying songs of each genre) and Lyrics Genius (for obtaining lyric data) APIs.

As shown in Table 1, each row lists a song’s metadata (name, artist, year, and genre), and lyrics. The data came in a pre-processed format with songs not in English, text indicating the section of the music and/or the singer of the section, and very common English words (that are unlikely to be indicative of genre) removed. The remaining words were lemmatized with WordNet, and are given as a single string in each row.



	artist_name	track_name	release_date	genre	lyrics	len
0	mukesh	mohabbat bhi jhoothi	1950	pop	hold time feel break feel untrue convince speak voice tear try hold hurt tr	95
4	frankie laine	i believe	1950	pop	believe drop rain fall grow believe darkest night candle glow believe go a	51
6	johnnie ray	cry	1950	pop	sweetheart send letter goodbye secret feel better wake dream think real	24
10	p√©rez prado	patricia	1950	pop	kiss lips want stroll charm mambo chacha meringue heaven arm japan b	54
12	giorgos papadopoulos	apopse eida oneiro	1950	pop	till darling till matter know till dream live apart know hearts till world fre	48
14	perry como	round and round (with mitchell ayres	1950	pop	convoy light dead ahead merchantmen trump diesels hammer oily kill gri	98
15	freestyle	opm medley: when i met you	1950	pop	piece mindin world knowin life come bring give world know give reason f	179
17	johnny mathis	it's not for me to say	1950	pop	care moment hold fast press lips dream heaven speak share glow grow p	21
20	st√©lios kazantz√≠dis	klapse me mana klapse me	1950	pop	lonely night surround power read mind hour night kiss lips hold tight une	30
23	st√©lios kazantz√≠dis	finito la mouzika	1950	pop	tear heart seat stay awhile tear heart game steal glimpse eye stare awhi	61

Table 1: Preview of Dataset

The original data also contains audio features drawn from the *spotipy* API, as well as scores for a topic model the researchers used in their study. However, for the purpose of this project, I extracted only the lyrics and genre from this data to perform classification on.

The distribution of genres is quite imbalanced, as shown in Figure 1 below:

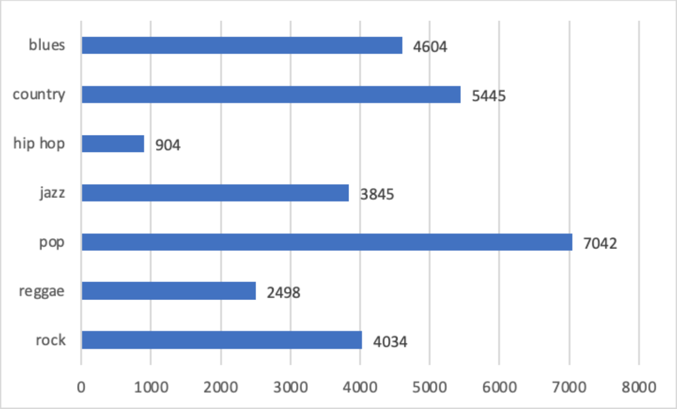


Figure 1: Distribution of Genres in Data

Methods

After reading in the lyrics-genre data, I first created a document-term matrix using a bag-of-words model computing the TF-IDF scores for each word. Only unigrams are considered in the final model - including bigrams in testing caused a significant drop in performance, likely due to the pre-preprocessing of the data; most consecutive words in the preprocessed lyrics would not be found in an actual sentence of a song.

In initial testing, I used various classification algorithms in the scikit-learn library, including multinomial naive Bayes, logistic regression, support vector machine, and random forest. Of these, logistic regression yielded the best balance of performance and runtime.

Seeing as the distribution of genres in the dataset was heavily imbalanced (see Figure 1), I expected the classifier to favor the majority genres (especially pop, which comprised 24.82% of the data). To correct for this imbalance, I evaluated the performance of the classifier in three cases:

1. Model trained on all songs, with genres evenly weighted
2. Model trained on all songs, with genre weights inversely proportional to the relative numbers of songs per genre
3. Model trained on a random sample containing 900 songs per genre, with genres evenly weighted

In all cases, a 90%-10% train-test split was used.

Finally, to get a sense of how certain words affect classification, I built word clouds showing the words in each genre with the highest tf-idf scores, and extracted the most informative words (those with the highest coefficients) for each genre in the logistic regression model.

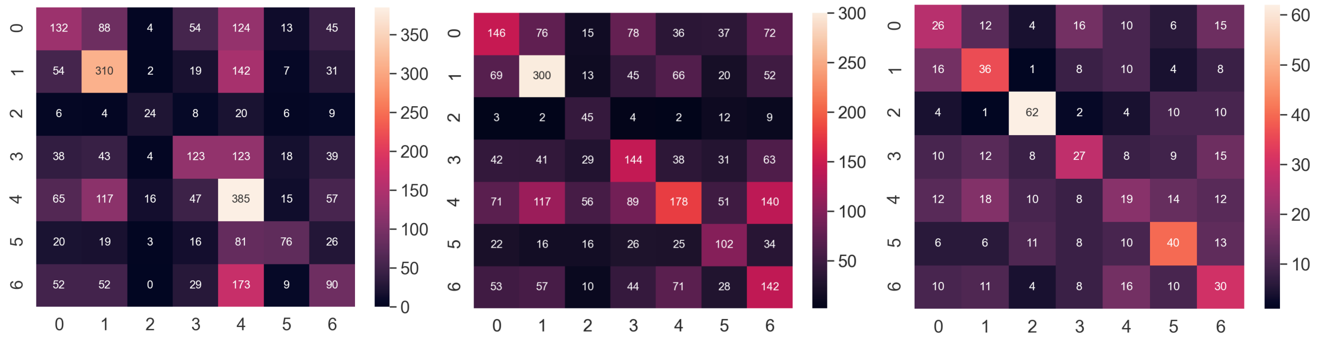
Results

Figure 2 shows the confusion matrices for each of the three cases described above, with a 90%-10% train-test split. Genres are labeled numerically: 0=blues, 1=country, 2=hip hop, 3=jazz, 4=pop, 5=reggae, and 6=rock.

Table 2 shows the F1-scores for each genre and the overall accuracy score in each case.

Word clouds displaying the words with the 50 highest TF-IDF scores (over the entire dataset) within each genre, are shown in Figure 3.

Table 3 shows the 15 most informative features for each genre (training the model on all songs with even weights).



All songs, even weights

All songs, balanced weights

900 songs/genre, even weights

Figure 2: Confusion Matrices for Logistic Regression Classifier

Genre	All songs, even weights	All songs, balanced weights	900 songs/genre, even weights
Blues	0.31923	0.33718	0.30058
Country	0.51753	0.51107	0.40223
Hip hop	0.36923	0.34483	0.64249
Jazz	0.35965	0.35208	0.3253
Pop	0.44	0.31843	0.22353
Reggae	0.39481	0.3908	0.42781
Rock	0.25641	0.30971	0.3125
Accuracy	**0.40169**	**0.37245**	**0.38095**

Table 2: F1-Scores by Genre and Overall Accuracy of Logistic Regression Classifier

Genre	Top 15 most informative words
Blues	baby, copyright, lord, woman, romance, blue, babe, women, mule, hound, mississippi, accuse, darling, mirage, guitar
Country	instrumental, heart, beer, lonesome, memory, whiskey, cowboy, truck, heartaches, tennessee, ries, texas, lay, wife, country
Hip Hop	niggaz, nigga, shit, rhyme, like, hiphop, dope, niggas, bitch, club, rappers, sayin, party, kick, hood
Jazz	spring, romance, interlude, repeat, tune, reality, sample, straighten, jazz, swing, strife, funky, vocals, lovely, funk
Pop	niggas, hoe, ohoh, sooner, flight, legs, feat, risk, club, imagination, kiss, plastic, split, kid, vibrations
Reggae	babylon, reggae, gyal, praise, africa, haffi, dont, waan, dread, dreadlocks, police, nation, people, chant, food
Rock	animal, television, crack, rape, monster, bleed, strangest, piss, away, tie, edge, death, brain, knees, fuck

Table 3: Most Informative Words in Logistic Regression Classifier by Genre

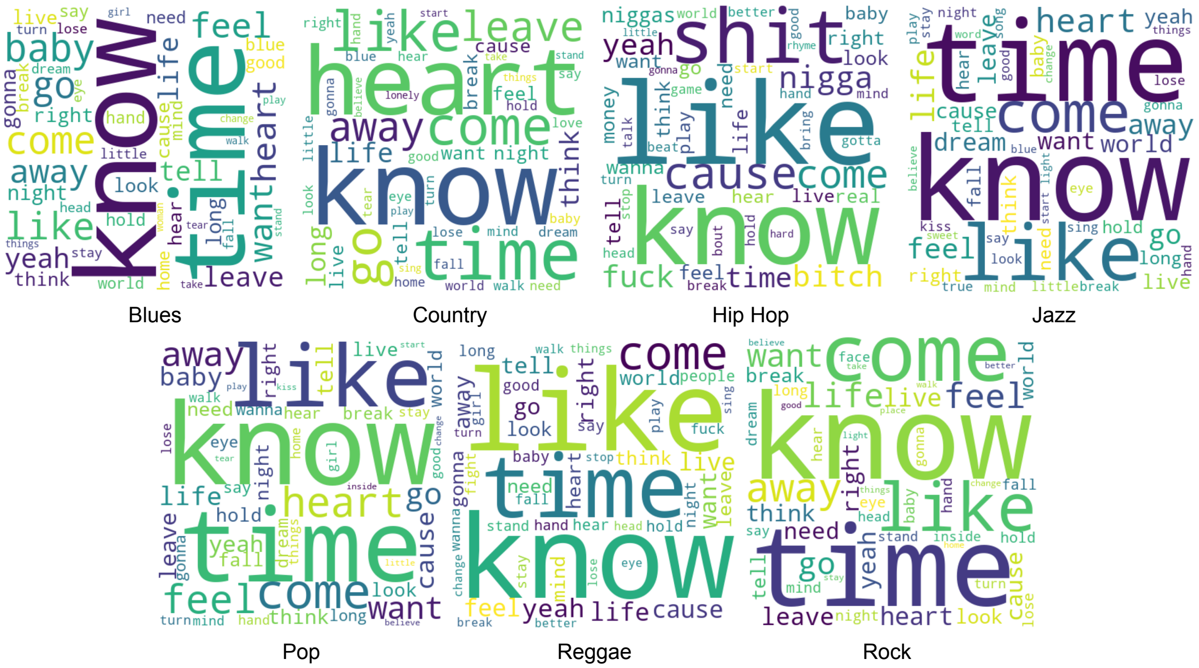


Figure 3: Wordclouds by Genre

Discussion

The best classification accuracy of 40.16% resulted from training the model on all songs in the dataset, without weighting any genres differently. While this is not necessarily ideal, it is quite reasonable - almost 3 times more accurate than random guessing. And in fact, it is not very far behind the 50% accuracy of Kovachev's most accurate model.

When considering all songs, hip hop had by far the lowest number of correct guesses compared to the other genres. Hip hop songs make up only 3.19% of the data, so it is understandable that the classifier would prioritize them lowest. However, when training on an equal number of songs from each genre, hip hop had by far the most correct guesses and a quite high F1-score of 64.249%. This is likely due to the extreme prevalence of profane language in hip hop: both the most common words within hip hop songs and the most informative words the classifier found contain a large number of these words. In comparison, the most common words in the other six genres are quite similar: "know," "time," "come," "like," etc.

Generally, the classifier seemed to perform best on country songs, with F1-scores above 40% in all cases. In particular, the country music genre has been frequently criticized for its overused lyrics: as a staff reporter from musictimes.com writes, "more than anything else, there is an undeniable trend that has taken the radio by storm: Trucks and booze." And indeed, these both appear among the most informative features: "beer," "whiskey," "truck." There are also a few American states ("tennessee," "texas") and words that seem to indicate struggling relationships ("lonesome," "memory," "heartaches," etc.)

In fact, there are many words among the most informative ones that encompass common themes, such as:

- Reggae: Cultural references ("babylon," "africa," "nation," "people"), slang words from non-English cultures ("gyal," "haffi," "waan," etc.)
- Jazz: Musical descriptions ("tune," "swing," "funky," etc.)
- Blues: Romantic descriptive words ("baby", "woman/women," "babe," "darling," etc.)
- Rock: Words depicting physical and mental suffering ("monster", "bleed," "death," etc.)

To further optimize this program in the future, I might consider adding additional features beyond a simple bag-of-words model (such as the length of a song or the number of unique words in a song) and fine-tune class weights to allow for more balanced accuracy of detection of all genres. Additionally, implementing a topic model is an intriguing option, especially when considering the common themes present among words the existing model identifies as important within each genre.

Conclusion

My goal was to develop an application to classify songs by genre based solely on their lyrical content. My working program makes use of a logistic-regression classifier and a bag-of-words TF-IDF model, and achieves accuracy of roughly 40 percent on a corpus of songs encompassing seven genres. In addition, I found that many of the words identified as most distinctive of each genre fall into themes that are commonly associated with these respective genres. Based on these results, I feel that developing a topic model would be a good next step to improve accuracy further.

References

“Country Music Stereotypes: Trucks, Girls, Whiskey & How Lyrics Have Changed Over the Past Decade.” *Music Times*, May 27, 2021 / 8:06 PM, 26 Jan. 2015, www.musictimes.com/articles/25560/20150123/country-music-stereotypes-whiskey-trucks-music-lyrics-changed-past-decade.htm.

Fell, Michael and C. Sporleder. “Lyrics-based Analysis and Classification of Music.” COLING (2014).

Kovachev, Dilyan. “How We Used NLTK and NLP to Predict a Song's Genre From Its Lyrics.” Medium, Towards Data Science, 2 Apr. 2019, towardsdatascience.com/how-we-used-nltk-and-nlp-to-predict-a-songs-genre-from-its-lyrics-54e338ded537.

Misael, Luan, Carlos Forster, Emanuel Fontelles, Vinicius Sampaio, and Mardônio França. "Temporal Analysis and Visualisation of Music." Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional, Evento Online, 2020. SBC, 2020, pp.507-518.

Moura, Luan; Fontelles, Emanuel; Sampaio, Vinicius; França, Mardônio (2020), “Music Dataset: Lyrics and Metadata from 1950 to 2019”, Mendeley Data, V3, doi: 10.17632/3t9vbwxgr5.3

Sala, Eduardo Muñoz. “Tokenization, Term-Document Matrix, TF-IDF and Text Classification.” GitHub, 17 Sept. 2020, github.com/edumunozsala/Intro-NLP-Text-Classification/blob/master/Intro_NLP_1_TFIDF_Text_Classification.ipynb.