Race and Music: Do Lyrical Themes Affect Mainstream Success?

Beyoncé’s 2024 album, *Cowboy Carter*, weaves together several musical genres, but specifically stakes a claim to country music as a black American artist. In 2016, Beyoncé, who is from the U.S. South, performed a country song at the Country Music Awards, and was met with backlash from predominantly white, Conservative fans who argued that, because she is black, she does not belong in the genre (Roberts). *Cowboy Carter* pushes against these ideas, but points to a larger issue of racial exclusion of black artists from mainstream chart success mediated through ideas of the racialization of music and genres.

While there are many different approaches used to predict which kinds of songs will lead to chart success, lyric analysis is an approach that has been relatively successful at predicting what kinds of songs will chart. In 2015, Singhi and Brown tried to predict which songs from a 2014 dataset would chart through lyric analysis, ignorant of their actual chart placement. Using algorithms that analyzed lyrical qualities like rhyme, meter, and stress patterns, they were able to “identify about half of the hits [or, songs that charted], while misclassifying only 12.8% of flops as hits” (Singhi and Brown 14). In 1982, Jaret analyzed what features led to successful country songs, finding that, especially for female artists, lyrical theme was “the single most important element” (Jaret 121) in determining the success of a record. While these analyses take different approaches, they both reveal a measurable connection between a song’s lyrics and its chart success. This project follows Singhi and Brown in their more quantitative, algorithmic approach, but builds on Jaret’s idea of lyrical themes as the subject of analysis. Introducing the additional lens of race, this project asks, “What lyrical themes have won black artists mainstream chart success over time?” Relatedly, it seeks to compare the lyrical themes of mainstream-charting songs from black artists with mainstream-charting songs from non-black artists, and with songs on “black” charts.

To answer these questions, I turn to *Billboard*, who constructs these charts by tracking how much a certain song or album has sold within specific genres. In *Major Labels*, Kelefa Sanneh traces the origins of Rhythm & Blues (R&B), and maps the racialized history of how the genre has historically been charted. The origins begin in 1942 when *Billboard* attempted to track music for “Negro Territories” (Sanneh 260), and created a chart specifically for black audiences, which has been know under different titles including, “Soul Singles” (Sanneh 259) and “Black Singles” (Sanneh 259), but is now known as “Hot R&B/Hip-Hop.” As a chart designed specifically to represent the success of black artists for black audiences, it serves as an interesting and relevant comparison to *Billboard*’s charts for mainstream audiences.

The first step in carrying out this project is constructing a dataset that incorporates both “black” and mainstream charts. In 2016, Kaylin Pavlik built a “50 Years of Pop Music” dataset, which aggregates a list of the top 100 *Billboard* ranked songs from every year between 1965-2015, and collects information about the song’s rank, title, artist, year, and lyrics. I adapt her methodologies to construct my own dataset which pulls the same information from *Billboard*’s Hot R&B/Hip-Hop Songs each year data is available (1942-2024). Since the data Pavlik uses and the Hot R&B/Hip-Hop chart measure slightly different sales, I will also construct a dataset of *Billboard*’s number-one singles chart to serve as a better foil for Hot R&B/Hip-Hop. In assuring that both charts measure the same trend, a potential dependent variable is removed, emphasizing the role of race in the analysis.

Pavlik’s method first begins with scraping chart rankings from Wikipedia using an algorithm she constructed in R (Pavlik). I apply the same methodology using the equivalent charts on Wikipedia, which have been verified against *Billboard*’s original charts. Next, I continue following Pavlik’s methodology, as she uses xml and RCurl to extract song and artist names from Wikipedia (Pavlik). Next, Pavlik scrapes song lyrics from websites that use systematized URLs , for example, she pulls lyrics from metrolyrics.com because their URLs are predictably structured as, “metrolyrics.come/SONG-NAME-lyrics-ARTIST-NAME.html” (Pavlik). She repeats this methodology with three sites: metrolyrics.com, songlyrics.com, and lyricsmode.com, leaving only 3.6% of her dataset without lyrics (Pavlik). Since it has been nearly a decade since she published this methodology, however, I expand her methodology through the inclusion of Genius.com, a lyric website which rebranded and began growing as Pavlik was constructing her dataset in 2016 (Headerbidding), so it would not have been as comprehensive of a lyric scraping site then as it is now. Genius also uses a systemized URL pattern: “genius.com/ARTIST FIRST NAME-ARTIST LAST NAME-SONG TITLE-lyrics,” making lyric scraping simple with the collected data. The inclusion of Genius also aids with the increased scope of my project, which starts in 1942, (compared with Pavlik’s that begins in 1965), as it has a large catalog of older songs. Further, Genius does not exclude instrumental music from the songs they collect, making it easier to keep track of whether there is lyric data left yet to uncover, or if data for a certain song does not exist.

I deviate again from Pavlik’s methodology to address some of the limitations of her dataset, and to address the unique needs of my dataset. First, after aggregating this data, I plan to migrate it from R to OpenRefine to clean the data, addressing Pavlik’s limitation of having multiple different spellings for the same word (i.e. “night” and “nite”). While there are some applications where a lack of standardization would be negligible, or even important to the research question, because my focus is on themes, making sure all word variations with the same meaning share standard spelling is important. OpenRefine can help standardize spelling variations, and correct spelling errors in scraped data. OpenRefine’s user guide explains how its cluster tool can identify matching text facets, and allow users to standardize spellings en masse (Rajput). Frank Donnelly puts this strategy into action in his effort to standardize a large quantity of government data. He uses OpenRefine’s cluster tool to identify inconsistencies in his data (like misspellings of “Brooklyn”), which he is then able to rectify with OpenRefine, saving him time in both identifying and standardizing inconsistencies (Donnelly 5).

Pavlik also notes inconsistencies with how data about songs with more than one band or artist is collected (Pavlik). To help remedy this data scraping limitation, an additional column can be created in OpenRefine to reconcile, link, or extend data from reliable sources (Ham). Wikibase is used to create Wikidata, an open source data collection that can be directly incorporated into OpenRefine, to reconcile existing data in one’s dataset (Delpeuch). In this case, a song title could be cross-checked against Wikibase, and, if an artist is missing from the dataset, that information could be expanded to include them.

Once all artists are identified, the last phase of constructing the dataset is to identify each of the artists’ race, specifically looking to classify artists into “black” and “non-black” categories. This would similarly mean reconciling information about each artist from Wikibase, and extracting that information into a new category within the dataset. Since race is more fluid and harder to describe than more objective categories like “song title” and “year released,” it is not data that can easily be scraped from Wikipedia with an algorithm like the ones Pavlik used to construct her dataset. However, this information does exist, and could be easily identified and applied with Wikibase in OpenRefine.

Now that the dataset has been created and cleaned, I use topic modeling, specifically Latent Dirichlet Allocation (LDA), to analyze all of the compiled lyrics, and pull out themes related to the research questions. The first step is to create a series of documents for the model from the song lyrics. jsLDA recommends that each document is no more than a few hundred words, and most songs are only a few hundred words (Minmo), so most of the process will be removing punctuation and capitalization from strings of text lyrics. For longer songs, however, part of this process will include splitting the song into a few documents. Each document is assigned an ID number followed by the text of the lyrics, and is sorted within a larger documents file with one document per line (Minmo). Additionally, a document of stop words like “the,” and “am” is uploaded so that more frequent words like those will not interfere with the model. This document is formatted with one word per line (Minmo).

I will run the model in jsLDA, selecting a quantity of topic outputs that can be increased or decreased based on the topics produced, and whether they are too narrow or too broad (Kulshrestha). To investigate racialized differences, I will divide the documents by which chart they originate from, and run separate models of each chart’s documents to identify any differences in topics. Since this dataset does not yet exist, it is difficult to determine how much information will be available, but if there are enough songs by black artists on the mainstream chart to run a model of just their songs, I will do that to compare topics that come from all mainstream artists to topics that come from just black artists on mainstream charts. If there is not enough data to run this model, I will run a model of all of the songs by non-black artists on the mainstream chart, seeing if there is a measurable difference in the topics produced without black artists on the chart. The whole dataset will be large, making it an ideal candidate for topic modeling, but, as I seek to compare trends between charts, some methodologies may have to be rethought to make sure there is still enough data to run a topic model.

I expect these models to produce outputs that largely overlap via broad themes like love and sorrow, but, given the large amount of raw data and the adjustable topic outputs, I would expect to be able to recognize some sort of racialized difference grounded in well-constructed categories. Sterckx ran an unsupervised LDA model on the “Million Songs Dataset,” and had people rank the strength of the topics on a scale from 1-3, with three representing a useful topic, and one representing a useless topic (Sterckx 50). Topics that identified strong lyrical themes were assigned a score of three, but other kinds of topics, like those linked strongly to a specific genre, or those that just rhymed or were cliché were assigned a one or two, based on their specific merits (Stercx 50). When the model was run with sixty topics, thirty-two were ranked at three, and only sixteen were at one (Stercx 50). With 120 topics, sixty-six were ranked three, and thirty-two were one (Stercx 50). With 200 topics, eighty-four topics were ranked three, and seventy-seven topics were ranked one (Stercx 50). While no amount of topics produced overwhelmingly positive results, at both sixty and 120 topics, more than half of what the unsupervised LDA sorted was ranked at three. This suggests that, though this methodology is not without flaws, and some broad or poor topics will likely appear in most results, with the right number of topics, there are well-developed topics that can be extrapolated to discuss larger trends. To that end, while it may be difficult to pull out extremely large, universally-applicable results, some racialized trends may emerge to contribute to an understanding of segregated music, and reveal potential lyrical factors in mainstream chart success.

This project could be used for a digital humanities approach to cultural studies, specifically African American Studies. Revealing which lyrical themes are the most effective for “black” (as defined by *Billboard*) artists could reveal interesting insights about how race factors into popular music. Are black artists given more room for vulnerability in songs popular on “black” charts versus songs on mainstream charts? Do different themes chart better on mainstream songs for black and white artists? How have these, and other, phenomena changed over time? While specific insights are hard to define without actually completing the project, this sort of research is important to the study of race, culture, and power.

I would also publish my dataset in Zenodo, and the code and methodologies in Github, giving credit to Pavlik for everything adapted from her work. Copyright laws do need to be considered in publishing the dataset, however, as it contains copyrighted material (song lyrics) that is comprehensive and untransformed, meaning that it may not fall under Fair Use exceptions to copyright law. The lyrics would be removed from the dataset, and replaced with an explanation of how to aggregate the lyrics.

Further, with an unanalyzed dataset or model output, it is a lot easier to misinterpret data and come to inaccurate conclusions. Considering the racialized aspect of this project, the risk of someone pulling out a piece of data inappropriately could mean that broadstroke arguments about black people are painted from poorly extrapolated data. While it is impossible to control how others will interact with published work, there is an ethical responsibility to try to minimize negative, inappropriate applications of data. To that end, I would publish a written statement alongside my methodology and dataset to encourage appropriate usage of the data, and minimize the chance of harmful applications.

This project seeks to identify where black artists have fit into mainstream American music charts compared to non-black artists on the same charts, and compared to “black” music charts through song lyrics. First, a dataset needs to be constructed, and include songs from mainstream and “black” charts over time combined with demographic data and scraped song lyrics. Once the dataset is created and cleaned in OpenRefine, different topic models are run in jsLDA to try to find racialized differences in lyric topics. These results can be translated into cultural studies research, but must align with copyright laws and consider questions of race before being published outside of a contextualized paper.

Works Cited

Delpeuch, Antonin. “Overview of Wikibase Support: OpenRefine.” OpenRefine RSS, 24 Nov. 2022, openrefine.org/docs/manual/wikibase/overview#:~:text=Wikibase%20is%20free%20software%20(a,and%20publish%20Linked%20Open%20Data.

Donnelly, Frank. “Processing Government Data: ZIP Codes, Python, and OpenRefine.” Code4lib Journal, no. 25, 21 July 2014, pp. 1–6.

“Fair Use - Office of General Counsel: Ung.” University of North Georgia, ung.edu/legal/copyright-compliance/fair-use.php#:~:text=The%20Fair%20Use%20doctrine%20provides,multiple%20copies%20for%20classroom%20use. Accessed 5 May 2024.

Ham, Kelli. “OpenRefine (version 2.5). http://openrefine.org. Free, Open-Source Tool for Cleaning and Transforming Data.” Journal of the Medical Library Association : JMLA, vol. 101, no. 3, July 2013, pp. 233–234, https://doi.org/10.3163/1536-5050.101.3.020.

Headerbidding. “Becoming the Genius.” Headerbidding, 8 Jan. 2024, headerbidding.co/becoming-the-genius/.

Jaret, Charles. “Hits or Just Heartaches: Characteristics of Successful and Unsuccessful Country Music Songs.” Popular Music and Society, vol. 8, no. 2, Jan. 1982, pp. 113–124, https://doi.org/10.1080/03007768208591184.

Kulshrestha, Ria. “Latent Dirichlet Allocation (LDA).” Medium, Towards Data Science, 28 Sept. 2020, towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2.

Minmo, David. jsLDA: In-Browser Topic Modeling, jsLDA, mimno.infosci.cornell.edu/jsLDA/. Accessed 5 May 2024.

Pavlik, Kaylin. “50 Years of Pop Music Lyrics.” GitHub, 2016, github.com/walkerkq/musiclyrics?tab=readme-ov-file.

Rajput, Aakash Amod. “Cell Editing: OpenRefine.” OpenRefine RSS, OpenRefine, 12 Jan. 2024, openrefine.org/docs/manual/cellediting#cluster-and-edit.

Roberts, Randall. “Conservative Country Music Fans Lash Out at CMA Performance by Beyoncé and the Dixie Chicks.” Los Angeles Times, Los Angeles Times, 4 Nov. 2016, www.latimes.com/entertainment/music/la-et-ms-conservative-cma-beyonce-dixie-chicks-20161103-htmlstory.html.

Sanneh, Kelefa. Major Labels: A History of Popular Music in Seven Genres. Penguin Books, 2022.

Singhi, Abhishek, and Daniel G. Brown. Can Song Lyrics Predict Hits?, University of Waterloo, cs.uwaterloo.ca/~browndg/CMMR15data/CMMR2015paper.pdf. Accessed 5 May 2024.

Stercx, Lucas. “Topic Detection in a Million Songs.” Universiteit Gent, 2012.

I pledge my honor that this assignment represents my own work in accordance with University regulations. - Alison Fortenberry /s/