Lyrical Analysis of Web Scraped Bandcamp Release Data

Sam Roberts-Baca & Thomas Tellner
University of Denver, COMP4447 Summer 2021

Github Project Link: https://github.com/samrubicon/COMP4447_FinalProject

EXECUTIVE SUMMARY

A new, unsigned music artist desires to uncover what similarities exist between music releases from various locations and genres in terms of their lyrical content. Using this information, he desires to know what similarities and differences exist in the lexicon of music released under various locations and genres. Accordingly, these results would guide him regarding what lyrical themes and topics are prevalent in various kinds of music. This project will be 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.

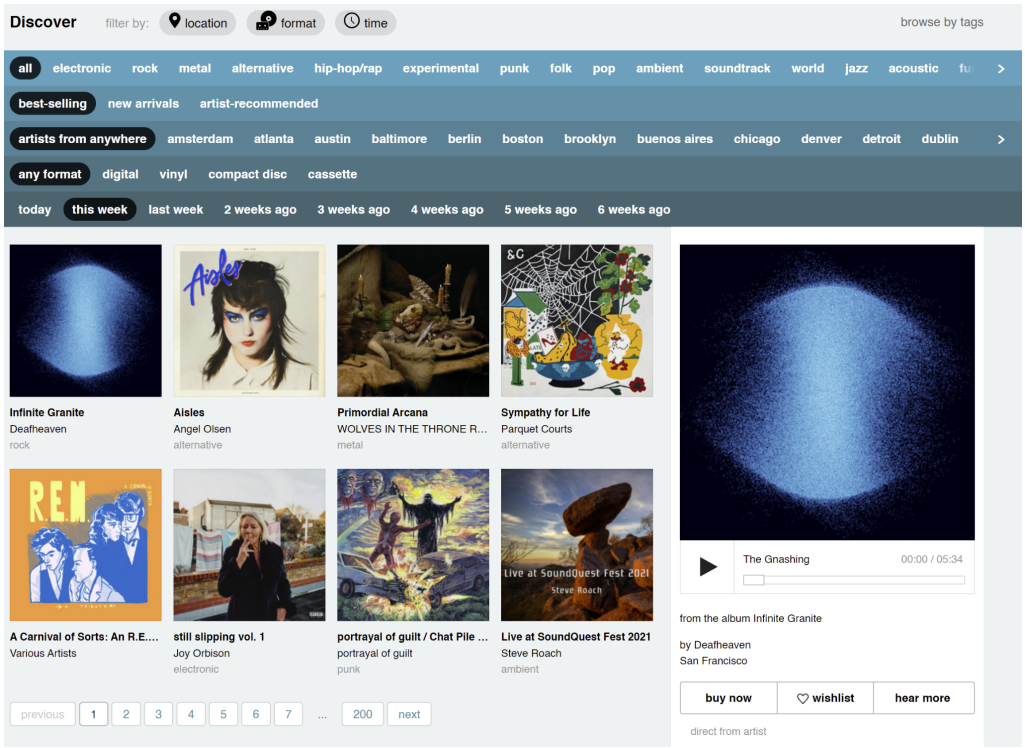

Figure 1. Bandcamp's Discovery Page

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 1118 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.

TASK DEFINITION / 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.

This study is less about commercialization than it is about trying to understand cultural similarities and differences present in broad topical categories. We are trying to examine various independently released albums being released in the year 2021. Lyrics from popular albums and new albums are grouped together by genre and location. We want to determine if songs released in different genres at this time in the English language share a common vocabulary and sentiment.

The inputs of our analysis are lyrics, release genre, and artist location associated with each release scraped from Bandcamp. The outputs of our analysis are tokenized lyric data, corresponding frequency tables associated with the tokenized lyrics, and word clouds and bar plots generated from the tokenized lyrics.

LITERATURE REVIEW

Hesmondhalgh et. al. [1] explored Bandcamp and Soundcloud as two audio distribution platforms, and argued that Bandcamp could be described as a more culturally democratic platform than Soundcloud in terms of the representation of artists on the platform.

Allington et. al. [2] focused on the cultural value uncovered in exploring music through locational analysis using platforms such as Soundcloud. This paper showed that artists on Soundcloud benefitted significantly in terms of their performance depending on their location, and that Soundcloud's design properties lead to emergent insulated musical communities of artists and fans largely enclosed by their location.

Our research is novel compared to past studies in that we are analyzing lyrical data obtained from Bandcamp in order to explore emergent cultural qualities present in music released by independent artists across various genres and lyrics on the platform.

QUALITY OF CLEANING

Data Cleaning

Web scraping is an inherently messy way of collecting data. In particular, there were a few specific challenges we faced with regards to scraping data from Bandcamp. We had initially assumed that all albums released in Bandcamp would be in U.S. standard english. However, many albums scraped were not released in English, and as such there were many special characters present in the collection of the scrape data. We had to make sure to save the data in UTF-8 format in order to prevent unrecognizable data that appeared to be corrupted when saving our data to CSV format.

Another issue was the fact that many lyrical scrapes appeared not to be written in standard English. Sometimes artists had names and lyrics that contained special characters, and this is reflective of evolving language use in a creative capacity.

There were a few fields where the actual information we wanted was contained within larger content blocks. We had to convert dates formatted as strings into datetime objects for example. Release titles and artist names were included in the same scraped string separated by a comma, and needed to be split.

Another field that was not included in this study was the number of fans associated with each album. The way to scrape these fans involved having our web scraping bot repeatedly click on a “more” button on the release page, to reveal 60 fans at a time on each click. The issue encountered is that there is no way to tell how many iterations our scraper needed to get to the true number of fans, so we set an initial limit of 1,000 maximum fans to collect. However, this proved to show that the majority of the best-selling albums released on bandcamp had at least 1000 fans. In the end, this field was not useful for determining a scale for which popularity could be meaningfully measured and compared, and was dropped for this study.

Handling Missing Values

The only field in our data that was consistently left blank was the *All Lyrics* column. This can be attributed to two possibilities: 1) some albums are instrumental and don’t have any lyrics associated with them, and 2) some albums that are not instrumental but don’t have lyrics listed. We dropped the releases that had no lyrics. The other field that had blanks was the Artist Location field. We decided to drop all releases that did not have an Artist Location listed.

Some albums have their release dates set to a date in the future, so not all of the tracks were listenable on their release page as of the date they were scraped. Therefore the *Track Duration* field for some tracks was unknown. We chose not to drop any releases with missing Track Durations.

VISUALIZATION

Additional Data Cleansing / New Feature / Attribute Creation

Once the data was chosen, located and scraped into a .csv file, and subsequently loaded into a pandas dataframe, work began on exploratory data analysis. Lyrics fall clearly in the category of “unstructured data” so the usual statistics - mean, variance, etc. - are of no use without some manipulation or transformation of the data. This was left until last. Before that, work was done on reviewing the additional columns that we captured in the scrape process, evaluating them for quality and either deciding to ignore them entirely for this analysis or perform some remedial data cleansing in order to make them more useful.

This work was more rudimentary than the work done while the data was being scraped - mentioned above in the discussion about missing values etc. The details of each attribute were listed above as well, for reference.

- “Release Genre” was immediately recognized as being a key attribute. Genres are as definitional to music - in not more so - than other categories such as location or time.
- “Release Sub-Genre” was not used but could be useful for later analysis to study why certain sub-genres were assigned and the differences in assignments to different music under the same main genre.
- “Search Format”, “Search Week”, and “Search Category” were technical specifications left from the search. They were included in order to confirm data pulled in the case of questionable results.
- “Release URL” was important; it allowed for the removal of duplicates.

- “Scrape Date”, “Release Title”, and “Artist Name” were not used at the level of analysis presented in this paper. However, it is easy to see how this analysis could be extended to using these fields, i.e. semantic analysis of two or more artists.
- “Artist Location” was instrumental in the analysis. Semantic analysis can be done on music of the same genre but from different locations. An example will be shown below.
- “Release Date”, “Tags”, “Track Info”, and “Number of Tracks” were also not used currently for varying reasons.
- “All Lyrics” was of course the main field without which this analysis would have been non-existent for obvious reasons.

In the end, the two most interesting questions that we wanted to address were well served by the fields we chose; “Release Genre”, “Artist Location”, and “All Lyrics”.

Before moving on to the analysis, however, a little more work was required. First, we used the “Release URL” to remove any duplicates that for any number of reasons were included in the scraped data. Secondly, we tokenized the lyrics and placed the resulting lists in a new column called, “Lyrics Tokens”. We used regular expressions to keep the contractions, such as “I’m” together as one word, remove newlines and remove common stopwords. Also, we found later that certain punctuation marks were different in one way or the other and therefore able to evade detection by the regular expressions or the code building the analysis. For example, the word “it’s” was counted in two separate instances because the apostrophe character was different. We replaced these with one common punctuation mark whenever we found this to be the case so it would be counted and totaled only once.

At this point, when it came to semantic analysis, questions about modifying the lyrics arose. Language and vocabulary are highly impacted by artistic license and can be a reflection of the mood or message of the artist. For example, while changing “can’t” to “cannot” may not necessarily add to the meaning of the lyrics, it can change the interpretation and be contradictory to the artist’s original intent. As a result, we did not expand or otherwise modify contractions, besides the change to normalize punctuation marks mentioned above. The results of the tokenization efforts, as a whole, were carefully evaluated to ensure that the integrity of the artists’ intent was not violated.

Data Visualization Methodology

The most revealing data visualization was driven by the questions we wanted to answer. Is there a significant difference in semantic content - specifically vocabulary - between genres? For two genres usually recognized as similar, is there a similarity in vocabulary used? This would be akin to asking, “What are the essential keywords used in a Country song? Home? America? Apple Pie? Do Folk songs contain the same messaging? How do Rap songs differ in vocabulary from Country songs?”

In order to provide an example analysis here, we chose “Alternative” and “Rock” as two similar genres. Then, we chose “Hip-Hop-Rap” as a genre considered very different in theory.

Second, are there significant locational differences in the vocabulary used? Or, for that matter, a difference in tone? For example, “Is this a difference in lyrics between New York City rappers and Los Angeles rappers?” This, if a difference exists, could be reflective of the rivalry that began years ago between “East Coast” and “West Coast” rap. In fact, many songs from that much earlier period would have directly contained references to each region, usually insults from one set made against the other or even threats of violence. Now, almost twenty years on, did that rivalry plant the seeds of different styles between the two?

The most obvious visualization to begin with would be word clouds. Clouds can be used to directly compare the content and frequency of words and concepts directly. In this case, they were very revealing.

In order to build the clouds, we wrote a straightforward loop that would iterate through the rows of the dataframe. When it found one of the three genres chosen by the user, it would add the token list to a larger list that would become that genre's corpus. For example, if the genre "Alternative" was chosen as the genre variable "gen1", all rows where "Release Genre" was equal to "Alternative" had the contents of "Lyrics Tokens" added to the list "comp1", which was the variable for that genre's corpus.

Before reviewing the visualizations, a few more notes should be made. We tried to 'variablize' the genres and locations in our code for future use. This means that we allow for easy modification to the analysis by users who want to do the same comparisons across other genres and locations. We felt that this would increase the value of the project for future use by adding that flexibility.

Data Visualization Results

Let's start with the first question: Are there differences in content between similar and different song genres?

First, we looked at "Alternative" and "Rock". Theoretically, being closely related in terms of tradition, lyrics should be very similar. We used the top 30 words occurring in each of the lyrical corpora that we were able to build. In Figure 2 we compare "Alternative" to "Rock":

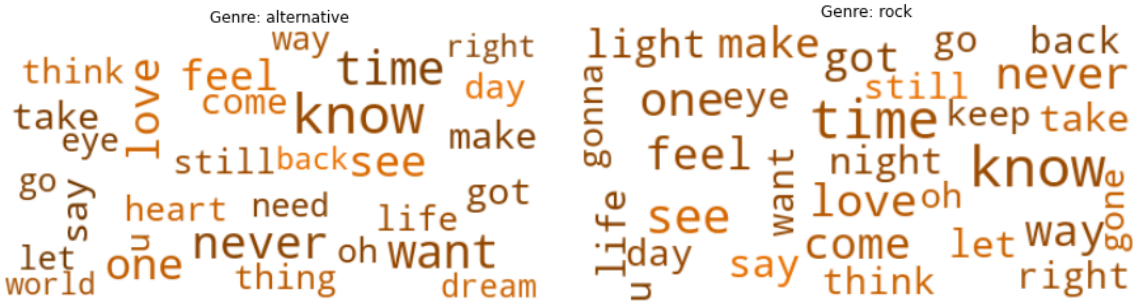

Figure 2. Alternative vs. Rock Genre WordCloud Comparison

Immediately apparent to the eye are the similarities: the predominance of words such as "know" and "time" in addition to "love", "never", and "feel". Overall, there are many words that occur in relatively similar frequency in both genres.

Now, we compare “Alternative” to “Hip Hop / Rap” as shown in Figure 3:

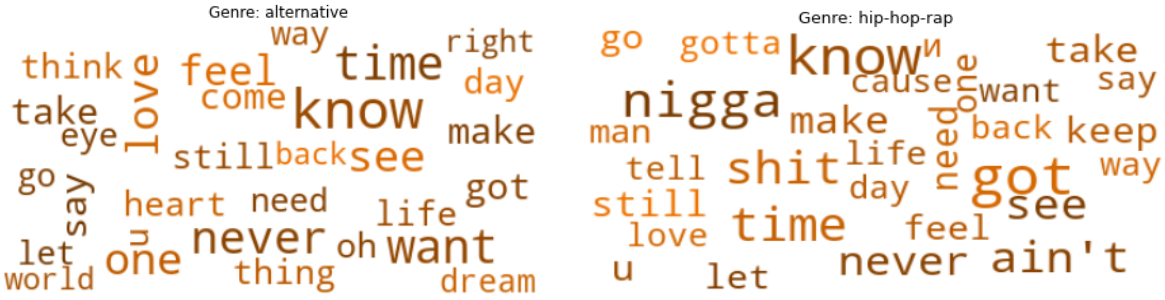

Figure 3. Alternative vs. Hip Hop / Rap Genre WordCloud Comparison

Once again, the result is pretty apparent - with one interesting exception - the vocabularies are different. The exception is “know” which occurs with similar frequency in both genres. Beyond that one word, the similarities break down, though there might be some connections in the form of the word, “life” and “love” though these seem to be common themes in any artistic medium.

Second, we compare one genre in different locations. In Figure 4, we observe the “Hip Hop / Rap” genres from New York City and Los Angeles in the same time period.

Figure 4. Hip Hop / Rap from New York, Los Angeles WordCloud Comparison

Again, similarities and differences are apparent. The similarities are there in some of the more frequent words. However, the differences are notable. For example, Los Angeles rap has much less use of the word “love” than New York. Also, New York seems to tend toward slightly more ‘vulgar’ words which creates an odd juxtaposition with the emphasis New York also places on “love”.

Of course, the observation of such a juxtaposition begs the question of context. For example, mention was made of a possible similarity in the frequency of the word “life” in all three genres. However, further but more complex analysis could reveal more information - especially when able to reveal more of the context. Using the word “life” in the sentence, “You are the light of my life!” versus the sentence “I am doing life for a crime I didn’t commit!” are two vastly different messages.

DISCUSSION & NEXT STEPS

Because of the nature of lyrics and artistic license songwriters can take with language, an obvious next step would be to build on this basic work and build an analytical package tailored to lyrics and the specific characteristics present in the analysis of songs - and as an extension poetry. Stopwords and contractions should be handled differently. Also, trying to expand the context of each word is also important as mentioned above.

Lastly, by building a record of common vocabulary and the relevant context paired with machine learning could also lend itself to the creation of machine-generated lyrics. This could take into account additional factors such as requested location, genre and topics. Additionally, machine-generated lyrics could incorporate current cultural events in a given location and consider how those events are portrayed in a given genre.

One conclusion from working with this data and reviewing possible questions and answers the data could provide is that there are numerous use cases for this type of analysis and expansion of it. From a marketing perspective, artists and managers could carefully review content by location or genre to make conclusions about popular content of songs in the artist's chosen genre. Indeed, the marketing applications are probably endless - especially with additional information. Bandcamp does not reveal much demographic information on fan profiles, which would be very valuable, but cross-referencing any information found on other sites might add to the value of the Bandcamp fan data. More readily available would be mapping fans to different genres to cross-sell artists and their work.

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.

REFERENCES

1. Hesmondhalgh D, Jones E, Rauh A. SoundCloud and Bandcamp as Alternative Music Platforms. Social Media + Society. October 2019. doi:10.1177/2056305119883429

2. Daniel Allington, Byron Dueck & Anna Jordanous (2015) Networks of value in electronic music: SoundCloud, London, and the importance of place, Cultural Trends, 24:3, 211-222, DOI: 10.1080/09548963.2015.1066073