More than Blue Jeans, Trucks, and Beer?
An exploration of country music lyrics

Ian Jones
Stanford University
ianjones@stanford.edu

Darby Schumacher
Stanford University
darbys@stanford.edu

Abstract

This research explores over one million lyrics from thousands of top-charting songs from 2013-2018, determining patterns in word usage not only in the country genre as a whole, but also between male and female artists within the country genre. Country music is compared to popular music, and a list of the "most country" songs is determined through a series of natural language processing techniques.

1 Introduction

In our project, we aimed to investigate the lyrics of country music in relation to the lyrics of other popular genres in order to find common themes in the country music genre. A main motivation behind this project was that we both felt that when we listen to country music we hear certain themes repeated over and over and over: love for God, love for America, driving trucks, country roads, falling in love, wearing blue jeans, etc.

We wanted to take on the challenge of both creating our own dataset from a variety of sources and analyzing these large corpora to tease out interesting insights in the country music genre. Using data and visualization techniques, we wanted to identify words that occur more often in country music than in an amalgamation of other popular genres, and conclude what words are the "most country" and if they align with these themes that we feel are common. We were inspired by a piece from The Pudding about lyrics in hip-hop music. As part of this project, we scraped data from Billboard and Genius to get the charts and lyrics, respectively, and then used D3 to tell a compelling story through "scrollytelling" that captures the users' attention and highlights our important findings. This project serves as the first stepping stone in a series of data projects around country music lyric analysis that ultimately could produce a model that could write its own country music lyrics that follow the same trends of the country music genre of today.

2 Related Work

We have found several reports that investigate lyrics across different music genres. We were originally inspired by a piece from The Pudding on hip-hop lyrics, and this is the piece of work most similar to what we are going to be doing. The Pudding article explores the most popular lyrics in hip-hop and explores the similarity between different artists based on their most used words. The style of data visualization in The Pudding was also a source of inspiration for us, as we wanted to showcase our insights in the "scrollytelling" format, which is how The Pudding wrote this article and many other articles. We want to do similar analysis with country music, but would like to explore more demographic trends between different country artists, especially around gender of artists.

The data analysis that inspired The Pudding article that inspired us, is an in depth look at words that are the most "metal" based on occurrence in metal lyrics. An interesting part of this analysis is that the metal corpus was compared to the Brown corpus, a famous corpus in natural language processing that is from the 1960s. We think this was a very interesting choice, because the corpus does not contain any lyrics, and is compiled from texts that are more than 50 years old. This seems like a rather large flaw in the analysis because what could have come from this analysis is not necessarily the words that are the most metal, but rather the words that are the most "lyrica" and could be genre agnostic. The Pudding significantly improved upon this in their hip-hop lyric analysis, by analyzing Billboard charting tracks (minus any song classified as hip-hop) and using this as the corpus of comparison. This is much a stronger analysis, since it removes the potentially confounding issue of comparing lyrics to non-lyrics. In our analysis, we compared the country music lyrics corpus to the Billboard Hot 100 corpus, to ensure that we were only investigating genre differences.

We felt that related work on music lyric analysis was missing an important exploration of demographic factors like age, gender, or country or origin, so in our analysis, we explored lyric frequency difference between males and females in country music.

To determine the "country-ness" of a given word, we compared each words frequency in the country music corpus (per 10,000 lyrics) over the words frequency

in the general music corpus. To determine what song is the "most country", we used cosine similarity. This is essentially a computation of the "closeness" of two vectors, so in computing the cosine similarity between every song vector with the "country-ness" vector, we were able to determine the most country country song and the most country non-country song.

In terms of tooling, we both found there was a significant amount of documentation online for scraping Billboard charts, scraping Genius lyrics, and implementing intricate interactive visualizations. We reached out to Matt Daniels from The Pudding who wrote the hip-hop story when we began this journey, and he guided us towards a few packages and libraries. We used Billboard Charts, a python library to scrape Billboard charts. We used GeniusR, which is an R package that generates URLs for song lyrics based on the artist and song title, and we used d3.js for all of our visualizations and embedded article text.

3 Methods

In this project, there were three main tasks. The first was data wrangling and data cleaning. The second was processing the data to produce the metrics that were central to our investigation. The third was building the visualization for web and writing a complementary embedded article to highlight key findings.

In the first task, we experimented with a few different tools to scrape the Billboard Chart data, knowing that we wanted to get all of the songs that had charted on the Billboard Hot 100 and the Billboard Hot Country charts from the past five years. We first used Beautiful Soup 4 and wrote our own Python script to scrape the artist and song title, but then we stumbled upon Billboard Charts, a python package that was made specifically for scraping Billboard charts. We ended up using Billboard Charts to scrape both the country and hot charts. We then began using the Genius API to get the song lyrics, but it turns out that Genius pays a very small licensing fee to record labels for posting the lyrics, so Genius redacts the lyrics from the API, so we had to look for another solution. After reaching out to Matt Daniels at The Pudding, he recommended that we look into GeniusR, which is an R package made for scraping lyrics by lifting them from the HMTL. In order to use GeniusR, we had to do a significant amount of string manipulation and regexes to parse certain patterns in the artist and title pairs that we had gotten from Billboard. We also removed any song that charter on Hot Country from the Hot 100 dataset. We wrote an R script that fixed any string issues with the title and artist (parentheticals, multiple artists separated by a comma, removing The Voice covers, etc), retrieved the lyrics using GeniusR, and then created a corpus for both country music and hot music.

The second task we undertook was data processing. For each word in each corpus, we calculated its occurrence per 10000 words, and filtered any song that was in a corpus less than three times. Since just looking at occurence in the country corpus doesnt necessarily tell the story of a lyric being extremely country in nature rather than just being a commonly used lyric in music, we calculated a lyrics country-ness by this equation: occurrence per 1000 words in the country corpus / occurrence per 1000 words in the hot corpus. A large number implied that the odds that the lyric is in country music is extremely high, and a low number implied that a song was really unusual in country music. By creating the country-ness vector, we were able to calculate cosine similarity between the country-ness vector and a given song for every song in the Hot Country dataset. The goal of this was to find which song was the most country, since a higher cosine similarity here implied that the vectors are most similar. We also repeated this same process with the Hot 100 data set, less the songs that also charted on the Hot Country charts, in order to determine the most country non-country song.

The third task we completed was the data visualization and seamless integration of text and visualization to tell a compelling story. We first diagrammed the visualizations that we wanted in the article in order to determine what outputs we needed to have from our data processing step, and then revised those scripts as needed. After outlining the visualizations we needed, we built frameworks for the visualizations with temporary data with far fewer data points, and then as the data was processed, added the real data points to already formatted charts. This allowed us to work concurrently without being each others bottlenecks. The visualizations depended largely on "scrollytelling", which we implemented to cause a series of animations that helped tell our story as the user scrolls through the article. Utilizing the Javascript D3 library, we created lineplots and scatterplots that allow in-depth interactivity, including searching and tooltip popups that give more information to the user. Furthermore, the lists of "top words" were initially rendered using HTML and CSS, however have added elements of Javascript that give more functionality to them in the form of on-hover events (highlighting words in the graphs).

4 Results

We presented our conclusions in the form of a web app, which can be viewed and interacted with at `https://ianjones763.github.io/countryMusicLyricsExplainer/` (It is recommended to view in Google Chrome).

The web app takes the form of a "scrollytelling" data visualization article which guides the user through an exploration of the dataset and our analysis.

4.1 Word occurrence in the country corpus

The article begins with a lineplot depicting the normalized occurrence (per 10,000 words) of several words in the country corpus.

This visualization was implemented with D3.js on

a logarithmic scale. The choice to plot the lineplots and scatterplots on a logarithmic scale is explained in Section 4.2. It is interesting to note that for each of the analyses conducted, the focus is on the comparison between the two corpora in question ("country" corpus and "other genres" corpus in the first section, "male country" corpus and "female country" corpus in the second section.

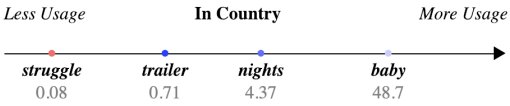

Figure 1: Lineplot depicting the normalized occurrence of several words in the country corpus

Therefore, although it is true that the word "baby" has the highest normalized country occurrence in the dataset, it is not necessarily the "most country" out of the four words in Figure 1 (and in fact, it is evident after comparison with the "other genres" corpus that "trailer" is actually the "most country" of these words). This is because the calculations regarding "most country-ness" depend on determining the odds that a given word appears in the "country" corpus or the "other genres" corpus.

4.2 The "Most Country" words

After plotting the occurrences of words in the "country" corpus, we did the same for the "other genres" corpus, and created a scatterplot plotting the "other genres" occurrence vs. the "country" occurrence of all the lyrics that we had scraped from Genius Lyrics. Words that appeared less than 3 times in both corpora were ommitted from the analysis (words that appear that infrequently have the tendency to be highly skewed towards appearing in one corpus or the other, and words that appear infrequently in both corpora are ommitted for our analysis). This initial graph was plotted on a linear scale, however this initial plot was very crowded and did not lend itself to viewing the trends very well, so we decided to switch from a linear to a logarithmic scale.

For this graph, word occurrence was plotted on a logarithmic scale. We decided to do this in order to achieve a more even spread throughout the body of the graph in order for the viewer to be able to distinguish individual data points in the plot. Normalized word occurrence in each dataset ranges from close to 0 to over 350, however the vast majority of words' normalized occurrences are less than 10. Therefore, a logarithmic scale spreads out the plotted points, lending to more readability. Furthermore, we care more about the comparison between the x and y values of each point, not necessarily the absolute value of each coordinate itself; this way of viewing the graphs allows for easier comparison as well.

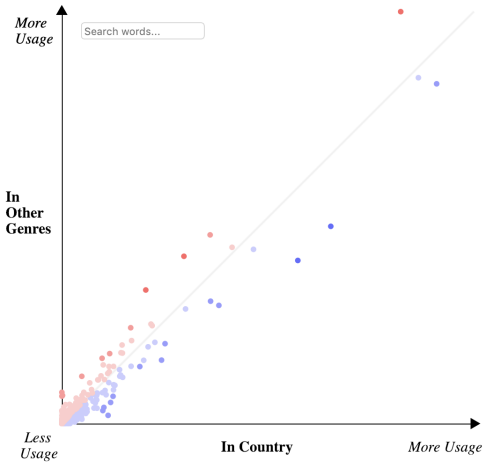

Figure 2: Scatterplot depicting the occurrences of words in the "other genres" corpus vs. in the "country" corpus, plotted on a linear scale

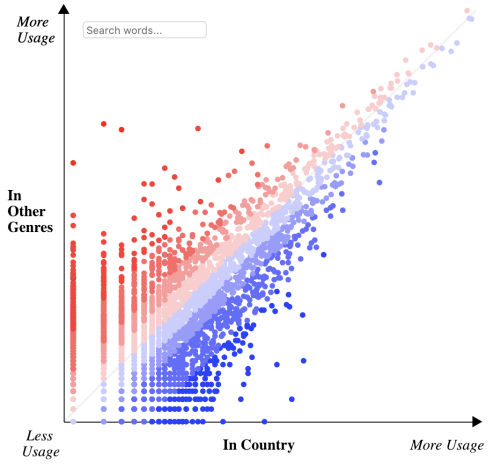

Figure 3: Scatterplot depicting the occurrences of words in the "other genres" corpus vs. in the "country" corpus, plotted on a logarithmic scale

Points are plotted on 3 dimensions: the x coordinate corresponds to the word's normalized occurrence in the "country" corpus, the y coordinate corresponds to the word's normalized occurrence in the "other genres" corpus, and the color of the plotted point corresponds to the odds that it appears in a given corpus: words that are more common in the "country" corpus than in the "other genres" corpus are plotted in blue on a gradient, getting darker as the odds that they appear in the "country" corpus is higher; words that are more common in the "other genres" corpus than in the "country" corpus are plotted in red on a gradient, getting darker as the odds that they appear in the "other genres" corpus is higher.

The user can explore the visualization in two ways: (1) hovering over a given point causes a tooltip with

more information about the word to appear, and (2) searching a word causes an animation to trigger, drawing attention to the point's location on the graph and causing a tooltip for the word to appear.

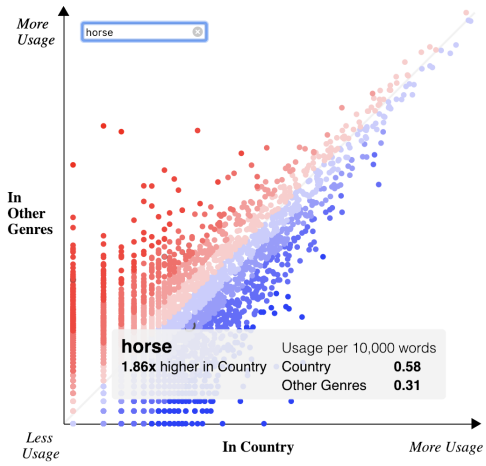

Figure 4: Visualization including a tooltip rendered in response to user query.

Words that hug the x axis have high "country-ness" (extremely high odds of occurring in the "country" corpus compared to the "other genres" corpus), and words that hug the y axis have low "country-ness" (extremely low odds of occurring in the "country" corpus compared to the "other genres" corpus). Words close to the grey diagonal (those that are much lighter in color) are used relatively evenly between both corpora.

The "Least Country" words

	Odds not in country	Country word count	Other word count
1. b***h	353:1	2	2631
2. ayy	291:1	1	1084
3. f**k	207:1	3	2309
4. ho	88:1	1	328
5. trap	70:1	1	260
6. bounce	47:1	1	174
7. mi	44:1	1	162
8. plug	42:1	1	157
9. boom	41:1	1	151
10. vibe	40:1	1	149

Figure 5: The "least country" words (lowest odds of occurring in the "country" corpus).

After plotting the points, we determined what the 10 "least country" and 10 "most country" words are in the dataset. Interestingly, the "least country" words are all words that appear very frequently in hip hop music. This seems to imply that the hip hop and country genres are diametrically opposite in terms of word usage in lyrics.

The "Most Country" words

	Odds in country	Country word count	Other word count
1. beer	106:1	143	5
2. southern	96:1	77	3
3. whiskey	58:1	187	12
4. homegrown	37:1	30	3
5. boots	35:1	102	11
6. hometown	34:1	45	5
7. Tennessee	26:1	48	7
8. gravel	23:1	19	3
9. sunset	22:1	53	9
10. Carolina	21:1	23	4

Figure 6: The "most country" words (highest odds of occurring in the "country" corpus).

The 10 "most country" lyrics are very much what we expected. It seems that the culture of country music is extremely strong, and the genre sticks to its roots (at least among top-charting songs). This could be for a variety of reasons, however we think that overall music genres distinguish themselves in two ways: musical patterns and lyrical patterns, and so it makes sense that country as a genre is so easily identified by word usage in songs.

Interestingly, the two words in the upper right hand corner of the graph (corresponding to the two words with the most overall usage in both corpora) are the words "I" and "you". It seems that songs in general are very personal and, at least in songs that hit the top charts, lyricists write about personal relations extensively.

In the web app, both lists allow interactivity as well: hovering over a word in the list highlights it on the scatterplot and renders a tooltip for it, similar to the response from searching or hovering over a word. This allows for the user to see that words on the "Least Country" list hug the y axis and words on the "Most Country" list hug the x axis.

4.3 Gender within country songs

We decided to further explore the "country" corpus, in particular we wanted to determine whether word usage in lyrics differed between male and female artists. We realize that this binary analysis of gender made a lot of assumptions about the genders of artists and gender identity in general, however in order to perform this analysis, we decided to split the country dataset based on the gender that country artists *present*.

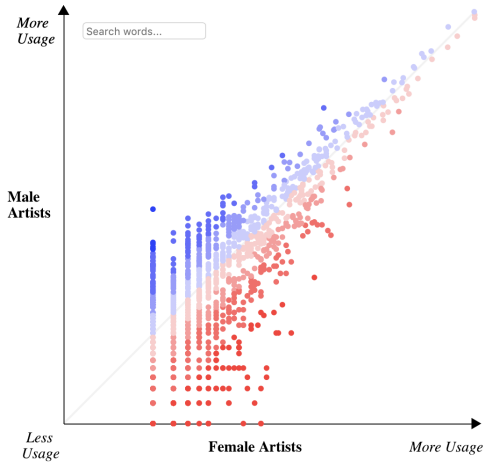

Figure 7: Scatterplot depicting the occurrence of words in the "male country artists" corpus vs. in the "female country artists" corpus, plotted on a logarithmic scale.

The "Most Male" country words

	Odds in male	Male word count	Female word count
1. beer	12:1	139	2
2. blame	6:1	72	2
3. body	6:1	70	2
4. turned	6:1	66	2
5. gettin	5:1	64	2
6. loud	5:1	62	2
7. tshirt	5:1	58	2
8. top	5:1	116	4
9. ones	5:1	87	3
10. taste	5:1	55	2

Figure 8: The words used most in songs by male country artists (highest odds of occurring in the "male country artist" corpus compared to the "female country artist" corpus)

After dividing the dataset based on male/female gender, we plotted the words again with the same techniques used for the "country" and "other genres" visualizations.

As can be seen from the graph, although there are many words that are used by female country artists but never by male country artists, the opposite does not seem to be true (This causes the large blank space on the left hand side of the graph). Furthermore, it must be noted that female artists were vastly outnumbered by male artists in the "country" corpus - only 14% of songs in the dataset are by female artists.

Similar to before, after plotting the "male country artist" and "female country artist" corpora, we determined the words that see the most usage in male country artists' songs compared to in female country artists' songs and the words that see the most usage in female country artists' songs compared to in male country artists' songs.

The "Most Female" country words

	Odds in Female	Female word count	Male word count
1. tux	50:1	17	2
2. runaway	35:1	17	2
3. horse	29:1	12	2
4. toy	25:1	15	3
5. mmmm	22:1	17	4
6. queens	19:1	19	5
7. boyfriend	19:1	13	4
8. smokin	18:1	19	6
9. biscuits	18:1	6	2
10. giddy	21:1	6	2

Figure 9: The words used most in songs by female country artists (highest odds of occurring in the "female country artist" corpus compared to the "male country artist" corpus)

Looking at the words used most frequently in male country artist' lyrics, we can see that the word "beer" is still on the top. Because there are so many songs by male artists in the "country" corpus, it makes sense that this would also be true in the earlier analysis.

It is interesting to see not only the difference between word usage in songs by male and female artists, but also the common words that they share. Like in the larger analysis, the words "I" and "you" are still the most frequently used words in both male and female artists' songs.

We analyzed all of the songs in both corpuses to find each song's cosine similarity with the "countryness" vector, which was a rating for each word in the combined corpus between the Hot 100 and Hot Country corpuses. In doing so, we determined the most country song that charted in the past five yeard on Hot Country and we also determined the most country song that wasn't country that charted on the Hot 100 (minus songs that also charted on Hot Country). Not surprisingly, the most country song was "Southern Style" by Darius Rucker, followed by "Power of Positive Drinkin"' by Chris Janson, adn then a slew of other songs each with Beer or Southern in the title. Interestingly, the top 21 most country songs are all performed by men, so this may feed into our hypothesis that we didn't see as many gender differences due to the small pool of female artists that chart on the Hot Country list.

The most country non-country song was a collaboration between Kygo and Selena Gomez, and in a strange convergence with our topic, features country melodies, even though it is an electronic dance track. The lyrics pick up on themes similar to country music, and was definitely a track inspired by the genre of country, so naturally, it has the highest cosine similarity with the most country vector. Which non-country tracks are the least country? Well that would be Spanish-language songs, as well as electronic dance tracks with limited vocals and a few rap songs. A more in depth analysis can be done in the future on this relationship between country music and hip-hip, since from our analysis, we think these two genres are diametrically opposed in terms of content and word usage, which also has historical and societal implications that reach far outside of the music world.

5 Discussion

By the end of our analysis, it is clear that the stereotypes that we set out to disprove are in many ways true. The country genre is very male-dominated and sings about themes you'd expect: Beer, America, trucks, whiskey, and those good ole gravel roads of Tennessee. We think that the audience of our visualization really learned about the importance of digging deeper into numbers, since we highlighted the reasons why we can't just look at occurrence in country music to determine how "country" a lyric is. Another aspect of our work that we sought to highlight in our visualization was the importance of entering data analysis with an open mind and without trying to steer the results in a certain direction. At the beginning of this project, we were actually trying to disprove stereotypes about country music, but only ended up reinforcing many of our original gut instincts about the genre.

From a purely visualization standpoint, we think our visualizations really capture the future of storytelling by seamlessly combining text and plots to give both the support of words and a framework of a story, but allowing the data to speak for itself.

The code for the data analysis is available at `https://github.com/darbsies/country-music`.

The code for the website is available at `https://github.com/ianjones763/countryMusicLyricsExplainer`.

The web app can be viewed and interacted with at `https://ianjones763.github.io/countryMusicLyricsExplainer/`

6 Future Work

There are several different directions that this project could go in, with having this analysis as the base. A piece of this analysis that The Pudding did was incorporate all songs by an artist that charted on the Billboard chart to create more robust artist corpuses, so that TF-IDF between artists could be calculated to understand which artists were the most similar to one another in the country music genre, and also to determine which artist were the most country overall. We both feel that this work can serve as a stepping stone towards an interesting machine learning project to create a model that can write country music lyrics that are coherent and follow the same patterns as country music that is popular today.

Another interesting direction this project could go is to analyze how music lyrics have changed over time, and find trends in these changes, such as differences in slang and prominence of usage of certain words. This would require much more data, which is available but tedious to obtain, as noted above. We could also analyze more than just the top charts of a given time period. Although the top charts are indicative of pop culture and popular taste in music during that time period, they also exclude certain genres from the analysis (for example, top charts in the past 5 years probably don't include many classic rock songs, but probably include a large number of hip hop and pop songs). Exploring less popular genres would be interesting and we predict that we would find similar results to the analysis with country music, namely that genres are distinct for a reason, and that these distinctions would show in the analysis.

Further potential continuation of this project would take the form of continuing to analyze demographic factors like our exploration of gender. A limiting factor in this analysis was that only 14% of charting Billboard country songs were sung by one female or an all-female group. As discussed in class, we also would like to explore other non-lyric features of the music like beat pattern, repetition, song length, and melodic composition, but that was outside of the scope of this project.

Acknowledgments

Thank you to our professor, Maneesh Agrawala, and our great TAs in CS448B for providing valuable feedback and tips for better visualization techniques.

References

[1] Matt Daniels. 2014. *The Words That Are Most Hip Hop*. The Pudding. *The Pudding* https://pudding.cool/2017/09/hip-hop-words/

[2] M. Fell, C. Sproleder. *Lyrics-based analysis and classification of music*. http://www.aclweb.org/anthology/C14-1059

[3] Iain Barr. 2015. *Heavy Metal and Natural Language Processing - Part 1*. Disqus, Iain Barr. http://www.degeneratestate.org/posts/2016/Apr/20/heavy-metal-and-natural-language-processing-part-1/

[4] Manos Antoniou. 2018. *Text analytics & topic modelling on music genre song lyrics*. Medium.

https://towardsdatascience.com/text-analytics-topic-modelling-on-music-genres-song-lyrics-deb82c86caa2

[5] A. Bou-Rabee, K. Go, K. Mohan. *Classifying the subjective: determining genre of music from lyrics*. http://cs229.stanford.edu/proj2012/BourabeeGoMohan-ClassifyingTheSubjectiveDeterminingGenreOfMusicFromLyrics.pdf

[6] P. Christenson, S. Haan-Rietdijk, D. F. Robers. 2018. *What has America been singing about? Trends in themes in the U.S. top-40 songs: 1960-2010*. Sage Journals. https://journals.sagepub.com/doi/full/10.1177/0305735617748205

[7] *Billboard Hot 100 1958-2017*. data.world. https://data.world/kcmillersean/billboard-hot-100-1958-2017

[8] silkandsilicon. 2018. *Billboard 100 Genre Breakdown*. Silk & Silicon. http://www.silkandsilicon.com/billboard-genres

[9] Guoguo12. 2018. *billboard-charts*. Github repository. https://github.com/guoguo12/billboard-chartsGuoguo12