Trends in Rap Music

How race, gender, and era can influence the lyrics of artists

Jake Vitale

CS 585, Intro to NLP, Final Report

Note: My data can be found here: https://github.com/jvitale94/NLP_Data

Abstract

Rap music is perhaps the most eclectic, varied, and important musical genre of contemporary music. It has gone through drastic shifts in musical qualities and subject matter in its life, and has developed as a tool for enacting social change and promoting political awareness discussions in various ways. It has also developed a tendency to discuss violence and sex in abundance. But how and when did these progressions and shifts come to be? This research delves into how rap music has evolved by exploring various subject matters rappers write about, as well as analyzing differences in content between race, gender, and eras of rap, all using various NLP techniques.

Introduction

Rap music is not only one of my favorite forms of art, it is an important part of modern culture in the United States. It is one of the most popular music genres, with some of the most prominent voices in the public sphere belonging to it. It has been used as a tool for political outcry and social change/reform throughout generations. It also has come under heavy criticism for its profane subject matter and misogyny. However, the genre did not always have these traits. It used to be fun and light-hearted, with songs such as Rapper's Delight by Sugarhill Gang, which has a verse about not liking the chicken that is being served at a dinner party. When and how did rap music progress to include topics such as sex and politics, and what artists use these topics most frequently? I attempt to map out a progression of these topics in the history of rap, as well as examine them across race and gender dividing lines. I also look at artist similarity to see how these topics correlate between similar or dissimilar artists. Lastly, I compare vocabularies of different divisions of rap artists to see how certain sects of rappers use

words differently, and in doing so see if rap music evolves in waves, with artists of each time period exploring similar topics.

Related Work

There has already been substantial work done in using NLP to analyze music. Mostly it has been LDA's" by James Thompson provides many baseline approaches to analyzing lyrics throughout eras of music. He based his research on "The Evolution of Pop Music" by Mauch, MacCallum, Levy, and Leroi, which used audio signal processing to categorize music. Thompson was able to categorize rap music with 60% success rate using his model based on lyrics, showing that rap music is a fairly unique genre when it comes to words. "Analyzing Rap Lyrics" by Dave King uses many complex methods to analyze rap music and comes to many interesting conclusions, but does not analyze differences across race or gender lines. It only does computations from era to era and by lumping artists together in different sub genres based on categories in the lyrics. It heirs more towards a categorization of lyrics based on subject matter. There are also numerous lyric generator research projects geared towards hip hop lyrics. However, many research projects exist about classifying rap music, or using bag of words approaches to point out trends in the music. However, I wanted to go deeper into this analysis and find trends amongst subsets of rap artists, regardless of their subgenre in hip hop. Looking at the splits between female and male or black and white rappers seems to be absent, or underrepresented, in the research that is out there, so I expected to find some interesting results.

Data

I attempted to find a public data set that would suit my needs. I looked at different research papers' data sets but many of them were not specific to rap music, and would require much work to clean up and eliminate the songs I did not need. Many of these datasets also did not contain enough artists from diverse groups that I needed to do my research, nor did they include enough contemporary artists, such as Kendrick Lamar, that I wanted to include in my research. Building my own dataset allowed me to control the specific artists in my data set,

which was crucial to having enough artists in certain groups (e.g. female rappers are not very common, but I wanted to include enough female artists to have substantial data).

I built a webscraper using Python to obtain my data. I used the requests library to issue HTTP requests, and then the html library to parse the returned HTML. I used azlyrics.com for obtaining my lyrics for many reasons. First, it has an extensive collection of artists, and was only missing a few that I was looking for. I used songlyrics.com for the artists that were not on azlyrics.com. Second, azlyrics.com did not have duplicate lyrics for different versions of the same songs, such as remixes. Third, the lyrics were in consistent parts of the HTML, making it very easy to parse. And lastly, the URL's were very easy to construct algorithmically based on the artist and song name, unlike many other sites that contained random numbers in their URL's.

The only problem with azlyrics.com was they had a block on their website that prevented mass amounts of requests from the same IP address. After about 300 requests, my IP would get blacklisted. I drew from a list of user agents randomly for each request, and put a pause in the program after each request, both of which did slow down the time between bans, but there was nothing I could do to prevent one IP from getting banned in the end. To get around this problem, I used the VPN at my school, which spat out a new IP address, from a pool of about thirty IP's, every time I connected to it. Eventually all of these would get banned as well, but the bans would always be lifted after a week or two, and I could run through all the IP's again. I looked into more sophisticated solutions, such as tor or using a cluster of servers, but my solution proved to work fine.

As for the actual data that I accumulated, I did research on the most influential and best selling artists of each decade, and chose accordingly. I did my best to include iconic artists, such as 2pac and Beastie Boys, as well as important artists that are a little less well known, such as El-P and Killer Mike. I also needed to make an effort to include artists that were female and/or white, which proved to be somewhat challenging. Since rap is dominated by black men, it was difficult to find influential white artists, as well as influential female artists. They exist (e.g. Eminem and Lauryn Hill respectively), but it was difficult to find enough that constituted significant data. I came to include artists that were the biggest in their respective sects of hip

hop (i.e. “who were the best white rappers?” instead of “of the greatest rappers I will include all of the white rappers”). This forced me to include artists that may not have the same level of clout that other artists in the data set have, such as Machine Gun Kelly, but gave me enough artists in every category.

The last challenge with my data was in dividing it amongst race, gender, and era. Gender was easy, as there were no mixed gender groups in my data set. Race was a little tricky, as there were some artists that did not fit into either Black or White. For example, Run the Jewels is a duo consisting of a black man and a white man. I chose not to include them in the split between races, as I did with any artist that did not fit into this binary divide. Era proved to be the hardest was to divide the artists. I ended up dividing the artists by the decade in which they rose to prominence. This method was somewhat subjective, but I felt necessary since the alternative, which is putting each album in the year of its release date, defeats the purpose of my research. If I analyzed the music based on the specific year it came out, I would come into situations where albums like Hot Sauce Committee Part Two by The Beastie Boys and Section 8.0 by Kendrick Lamar would be placed in the same era, as they both were released in 2011. However, these albums are drastically different as Hot Sauce is the last album released by the Beastie Boys in a nearly thirty-year career, while Kendrick was just starting out his career in 2011. These artists are from very different eras of hip hop and have very different influences and outlooks on music. They both grew up in different cultures, music scenes, and overall societies, and I wanted to measure the impact those factors have on artists, so I decided to split up the data by a somewhat subjective measure.

Overall I had 90 artists in my data set, 11,332 songs, and 5,704,055 words. There are 61 men and 29 women artists, 75 black and 13 white artists (2 did not fit this divide), and 15 artists from the 1980’s, 30 artists from the 1990’s, 24 artists from the 2000’s, and 21 artists from the 2010’s.

Method

My research consisted of three main components. The first was using cosine similarity to determine which artists were the most similar/dissimilar to each other. My thinking was that

artists in similar categories would have a higher similarity score as artists in disparate categories (i.e. Dr. Dre and Snoop Dogg should be more similar to each other than Dr. Dre and Iggy Azalea). I was also curious which artists would cross those lines and have a high similarity to artists unlike themselves. For example, I expected Eminem and Tyler the Creator to have a high similarity since they are both shock rappers, who say very vulgar things in their songs.

Second, I used the word2vec model to create word embeddings. The basis of the word2vec model is a skip gram for context and then principal component analysis for dimensionality reduction. The skip gram looked at the two preceding words and two following words, and the PCA reduced the embeddings to two dimensions. These final embeddings allowed me to find words in the vector space that are close to a target word, and therefore similar to a target word. Using this method, I can see how certain categories of artists use words and draw conclusions about topics in their songs. For example, using the word "drugs" as the target word I could see if a category of artists thinks of drugs in a positive light, or in a negative light based on its similar words.

Lastly, I ranked the artists based on certain topics. Using the word embeddings created by word2vec and then finding similar words as target words, I created vocabularies for three categories: vulgarity, politics, and sex. I looked at the neighboring words for words I knew would be included in each category, such as "America" in politics, and then used the neighboring words to build a list of words that represented that category (these lists are in the code, under the count_TOPIC methods). I then ranked the artists based on frequency of these words.

Results

Here are my results for artist similarity, both the top 10 most similar artists and the top 10 most dissimilar artists:

<u>10 Most Similar Artists</u>

Geto Boys and Scarface - 0.992933744525

Snoop Dogg and Dr. Dre - 0.992390429192

Ghostface Killah and Wutang Clan - 0.99167904162

Ghostface Killah and Raekwon - 0.991009898274

Jay-Z and Dr. Dre - 0.990954253426

Jay-Z and Snoop Dogg - 0.990792083987

Snoop Dogg and Ludacris - 0.990453297578

Kendrick Lamar and Kanye West - 0.990314479201

J Cole and Kanye West - 0.989592138566

Ice Cube and Geto Boys - 0.988979856623

10 Most Dissimilar Artists

Raekwon and House of Pain - 0.805045608405

Raekwon and Lil Mama - 0.804315263571

Beastie Boys and Wutang Clan - 0.803673363707

Raekwon and Khia - 0.799618629681

Wutang Clan and Monie Love - 0.798875009709

Shawnna and Wutang Clan - 0.783777841454

Khia and Aesop Rock - 0.765678682258

Raekwon and Beastie Boys - 0.762359653281

Monie Love and Raekwon - 0.75793193038

Raekwon and Shawnna - 0.753004619425

These results follow my hypthesis very closely – artists in the same categories are very similar, and artists in different categories are not similar. Of the 10 artists who are most similar, they are all black men and all of them are either in the same era of hip hop, or differ by one era (i.e. one is in 1980, the other in 1990). This difference in era can also be attributed partly to the subjectivity in placing artists in an era, as some artists maybe should have been included in the same era, but were serparated, such as Ghostface Killah (1990) and Wutang Clan (1980). And notably, all pairs except for one, Kendrick and Kanye, have at least one artist from the 1990's. The could be a product of the revolution of gangsta and political rap that emerged in the 1990's

and the homogeneity in topic choice for many artists. Furhtermore, as expected, the transitive property holds for similarity. Every permutation of Snoop Dogg, Dr. Dre, and Jay-Z exists in the top 10 and other such trios all lie close to the top of the list.

Conversely, dissimilar artists all lie in disparate categories, with one exception, that being Raekwon and House of Pain. This again points to the homogeneity of artits in the same categrories, and the differences in content amongst artists in different categories. Eras in rap seem to be very real, in that there are waves of artists that all become concerned with similar topics and feed off of each other. Additionally, in the least similar artists, almost all of the pairs involve a woman differing from a man. This highlights a potential divide in gender for rappers, as women and men seem to talk about disparate topics. Lastly, Raekwon is among the 6 most dissimilar pairs. I thought this may have been an error in my data, until I realized he is also in the fourth most similar pair, so his data could not be corrupt. One conclusion I can think to draw from this anomaly in Raekwon's similarities with other artists is that Raekwon is one of the most unique artists in rap, which I would not have thought before doing this research.

As for arrtists that differed in multiple categories who were still similar, the most similar artists who were also very different categorically were Mac Miller and Kanye West with a cosine similarity of 0.98851658949. This surprised me as Kanye has many diverse subjects in his songs, and is even political in many cases, while Mac Miller raps mostly about drugs and sex. Eminem and Jean Grae were the most similar artists to cross the gender divide with a cosine similarity of 0.987131244114. They differ in every category as Eminem is a white man from the 1990's (he could have been placed in any category really, but his first two albums came out in the 90's) and Jean Grae is a black woman from the 2000's. There does not seem to be much rhyme or reason to artists that cross over in categories. Some artists are just similar to others.

There were also many intersting results from creating word embeddings for different categories of artists, and looking at the surrounding words for certain target words. This method can be explored for thousands of target words, each word giving interesting results. For the sake of brevity, I will focus on five words – "drugs", "bitch", "money", "america", "black". I will not talk about every word that neighbors each target word, as some are insignificant, such as "um", but rather the neighbors that I think point to the most significant tredns.

The word “drugs” was fairly divisive between many groups. The split between Men and Women was most interesting, as Men had words such as “party”, “ballin”, “hardcore”, and “cash” associated with “drugs” highlighting the fun side of drugs. This was the case for many other groups, as common neighbors of “drugs” for other groups were words like “dreaming” or “warm”. However, for Women, drugs neighbored words such as “funky”, “buyin”, “projects”, and “critically”, highlighting the negative side of drugs – being relegated to the projects and not focusing on the warm feeling or parties. The split between how men and women see drugs was apparent. Additionally, for the 1980’s the word “drugs” was not used enough to even be included in the top 1,000 words, while for the 1990’s the word gained interesting meaning. It neighbored the words, “police”, “media”, and “pressure”. The word was immediately used in social or political contexts in the 1990’s, showing that this decade was the time that rap started to become a genre of social awareness and politics.

The word “bitch” had its most significant results when it was used by Women. Most other groups had neighbors that highlighted sex (“ass”, blowing”) or lavish lifestyles (“benz”, “drink”, “balling”) but women had words that highlighted companionship. The neighbors included “ho”, “gangsta”, and “n**ga”, the last of which is usually used as a term of endearment or friendship. The word “bitch” seems to have been distanced from its misogynistic origins when it is used by women, and instead it is used as a term for friends.

“Money” is one of the central topics of rap music. Wealth is glorified and boasted about commonly in rap music, but the results are not completely homogenous when the neighbors are looked at across artist categories. Men seem to focus on negative aspects of wealth, with words “evil”, “sin”, “disaster” being neighbors for “money”, possibly showing how men focus on the problems that come with having money in a dangerous world. Women on the other hand have a totally different take on wealth, as words “legs”, “thighs”, and “bounce” are neighbors, suggesting the superficial mindset that comes with wealth. The 1990’s and 2000’s focus on the danger that comes with wealth, with words “glock”, “casket”, “blow”, and “guns” used, suggesting the drugs and violence that come with wealth. No group focuses primarily on the positive aspects of wealth.

“America” is often billed as a land of opportunity, but rap music has spent decades attempting to debunk that myth. Most groups tend to have critical words associated with “America” with words “phony” and “secret” being used by Women, “stupid” and “wicked” used by white rappers, “chronic”, “trapped”, “government”, and “thugs” used by 1990’s rappers, and “hammer”, “broken” and “weapon” used by 2010’s rappers. These neighborings suggest a common theme amongst most groups of rappers of criticizing America.

Lastly, “black” had notable disparities in neighboring words. Men had a surprisingly negative word group neighboring “black” with words “pimp”, “violence”, “ass”, “f*cker”, “cut”, and “p*ssy” all being used. Given that most of the men in the dataset were black, I find it counter-intuitive that so many negative words are used with black. White rappers have a very interesting take on the word, as the words “class” and “culture” neighbor “black”, possibly showing an analytical approach to race in white rappers’ music. The 2000’s and 2010’s share this characteristic of neighbors, as 2000’s have the words “politics”, “slave”, “hell”, and “pride” neighboring “black”, and the 2010’s have “believe” and “class” neighboring. These results again highlight the social awareness that is prevalent in rap music, especially contemporary rap.

Finally, here are my results for the artists who use the most curses, use the most political words, and use the most sex words:

<u>Top 10 Cursers</u>

Khia 0.0567057390775

Tyler The Creator - 0.0414387411015

Gangsta Boo - 0.0385695732278

NWA - 0.0382653061224

Freddie Gibbs - 0.0373449615691

Geto Boys - 0.0331262586613

Earl Sweatshirt - 0.0303975058457

Young Jeezy - 0.0302008127195

50cent - 0.0301111314011

Lil Wanye - 0.0294253163993

Top 10 Political Artists

Lauryn Hill - 0.00829875518672

Lisa Lopes - 0.00744592079081

Public Enemy - 0.00734742452074

Mos Def - 0.00678227168836

M.I.A. - 0.00617122706197

Nas - 0.0057990328348

Jedi Mind Tricks - 0.0051087607264

Kanye West - 0.00499806453082

Salt N Peppa - 0.00493042203651

Kendrick Lamar - 0.00485125018028

Top 10 Artists Who Talk About Sex

Khia - 0.0395037296392

Gangsta Boo - 0.0262768431544

NWA - 0.0258335728658

Tyler The Creator - 0.0224803297115

Freddie Gibbs - 0.0207679386464

Trina - 0.0205196697426

Geto Boys - 0.0202751134928

Lil Wanye - 0.0192031866588

Danny Brown - 0.0176301452785

Azaelia Banks - 0.0170066566459

There are many notbale results here. First, the similarity in the cursers and artists who talk about sex is striking. 7 artists are on both lists, showing how the two topics use extremely

similar language. Additionally, four artists, as well as the top two, on the sex list are women, which I found notable. Men are generally thought of as being sex-crazed and painting women as nothing more then objects of desire in rap, but women appear to be just as sex-minded as men. Additionally, 7 of the 10 artists in the top sex list are from the decades 2000 or 2010. This shows that as rap has progressed, it has become more sex oriented and possibly more vulgar. This is consistent with my idea that rap used to be more light hearted and clean, and has become explicit as it has progressed through generations.

The top 10 cursers list is also indicative of a more recent trend in vulgarity as 7 of the 10 artists also are included in the decades 2000 or 2010. There are 8 men and 2 women on this list however, which may indicate that men curse more than women, however this just barely the case. Men curse more than women at only a 10% higher rate. Additionally, women talk about sex at a 9% higher rate. It may be fair to say that men and women have nearly no difference in vulgarity in their music, which I found very interesting.

Lastly, the top 10 political list has interesting implications, as 4 of the artists came from the decades 1990 or 1980 (Salt N Peppa was included in the 1980's but could have gone in the 1990's. I chose to include them in the 1980's because their first two albums came out in 1986 and 1988). Still, this shows that rap has been politically aware for decades and has been used a tool for social change throughout much of its history. The political list was also the only list of any top 10 list to include a white artist, Jedi Mind Tricks. However, the rate that white rappers talk about political topics is only 3% lower than black artists, which correlates with the finding in the word embeddings above.

Discussion and Further Works

Overall, this gave a very interesting overview of rap music from a statistical standpoint. There are many implications to be seen from the results given by the algorithms I ran, and all of them are important to understanding rap music and its place in society. Next steps would possibly include expanding the dataset and trying to find trends within a certain category, such as seeing how the rap of the 1990's progress as the decade went on. It would also be

interesting to see how neural network more specifically tailored to semantic analysis would impact drawing out meaning from songs.