COMP3517 Written Report

Abstract

Music, similar to other artistic works, offers a lens into the lives of the artists that produce these poetic experiences. Music can also reveal important details about a country's history and the experience of everyday people in this country. This report details an attempt at using various off the shelf NLP tools to analyse the change in language in popular American musical works and compare that to the greater context of the American historical experience between the years of 1890 to 2019. Due to the nature of the artists that tend to populate the mainstream music charts and in attempt to offer a diverse perspective, I also include popular African American artists and rap artists that tell unique stories that are less glamorised than mainstream pop songs. These artists were chosen based on their cultural effect on hip hop, soul, R&B or funk and their involvement in African American music culture.

1 Introduction

When we think of a cultural phenomenon the we think of celebrities and popular theatrical works. Movies like Titanic and celebrities like Leonardo DiCaprio are very influential on modern day pop culture. This is no different in the realm of music. Nowadays, the most popular and well known musical artists globally are (for the most part) American. Part of this is because English is the most widely spoken language but also because American culture has managed to spread to most parts of the world and so have the works of American artists and celebrities. These artists use their music to tell stories that, accompanied by enticing instrumentals, are able entertain millions and create massive followings. It does beg the question has popular music stayed the same lyrically? Talking lyrically or grammatically one may assume that the lyrics have changed just as much as how we speak has changed. However, written word and spoken word for the most part has not changed as much as you'd think. Meaning behind some words and the format of sentences may have changed. However, the core elements have remained the same as I hope to show is the same for music. The following sections detail the results obtained when analysing the changes in the lyrics using various off the shelf NLP tools. It is worth mentioning that a larger dataset could have been compiled for both lists used in the analysis of music lyrics in hind sight.

2 Background

2.1 NLP

NLP or "Natural language processing" is a field that began as a number of different standalone fields. The list of fields include, linguistics, machine learning, entity recognition, etc. The aim of NLP is to make use of different approaches in these different fields to teach a machine to extract meaning from text. NLP has come a long way from its inception.. Below is Fig. 1 which shows an image taken from Khurana et al. That shows the recent developments in the realm of NLP

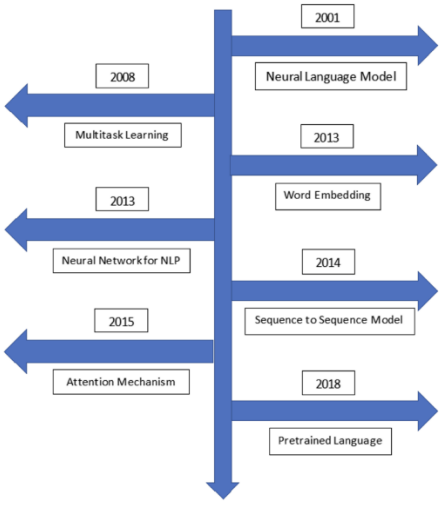

Figure. 1: Showing NLP developments from 2001 to 2018

These developments have made it easier and quicker to make use of NLP techniques such as the ones discussed in this project to automate extraction of meaning from bodies of text.

2.2 American history

America as one of the most influential countries in the world and has a very complicated history. From the importing of slaves from Africa brought by Dutch ships in the 1600s, to the abolishment of slavery and later introduction of Jim crow and segregation, to a country that prides itself on its diversity of ethnicities, it is fair to say that America has a long and complicated history. Despite this or because of this (depending which way you want to look at it). American music tends to be quite diverse. It has aspects of many different cultures and perspectives. If one is looking for the African-American perspective then you would look to soul, funk or hip-hop. If you're looking for the general/societal perspective usually you could look to the mainstream music which deals with many different themes (usually specific to the artist). If you're looking for a more Texan perspective then you might look at country music. Although the way I put it almost makes it seem like "black and white", in reality America is more than that. However, due to how long African-Americans have been in America and due to the way they were brought to America, their stories told and perspective which is reflected in their music tends to vary far more from what is portrayed in mainstream American music. Where popular mainstream music often talks about relatable concepts common to nearly all people (like hope or love), hip hop often describes elements of struggle, strife and a journey to success. Of course this also could be because I just do not know about Mexican/Latina/Spanish music as well as I do Hip Hop therefore I have a bias towards African American music/culture.

3 Methodology

In this section we discuss what methods were employed in order to be model the changes in language.

3.1 POS Tagging

POS (Part-Of-Speech) tagging is a useful tool for investigating the types of speech that are present in a body of text. In this context I am using it to try identify changes in grammatical tendencies in the different popular musical works. For POS tagging the function "process_music" shown in Figure 2 takes a number of arguments and returns the POS distribution for each of the decades or artists depending on arguments declared (for details about arguments passed to functions for intended output refer to the README.txt). In some cases it is easier to use this method for grammatical analysis. As this method allows observations to be made and assumptions be drawn from specific distributions by looking at the more commonly occurring POS.

3.2 Topic Modelling

In order to topic model I make use of gensim's LdaModel which models the probabililities of specific words appearing for the latent topics present in the dataset. These are topics that are not directly observable and are obtained by the model through analysis of the semantic structures present in the text. Topic modelling using the Gensim library is yet another way of observing what words are commonly used by artists in their music. It is useful for identifying relationships and emotions that humans may not usually be able to detect.

3.3 Sentiment Analysis

Sentiment analysis is a highly dependable technique that allows us to extract the general sentiment expressed within the lyrics. Using this method it will be possible to check if certain historical event should have correlated with an increase or decrease in specific sentiments. Sentiment analysis tends to be more dependable because artists rarely use words that express one sentiment to express another sentiment. It is also because the dimensionality of what the machine is trying to predict is significantly reduced when compared to the other techniques. I make use of the vaderSentiment module which has the capability to measure sentiment present in bodies of text. It is worth noting that, although sentiment analysis is not flawless. It works well on general text however, in cases where literary devices like metaphors are often used and the representations of meaning are hidden in the words there is a possibility of incorrect scores being given. There is also a possibility of bias towards certain words, as later discussed, interfering with the sores.

4 Data

4.1 Collecting data

In order to collect the data for this investigation it was necessary to first decide on the what data would be appropriate to include. Since there wasn't an already existing corpus of data that would fit this specific task creating an appropriate corpus was necessary. Since there were only 130 songs that had to be included on the side of songs by decade (technically 128 excluding the non English ones) this was done manually as it would take less time to do it this way than create a script to automate the process. Furthermore, implementing web scraping would have taken time to learn and so it would have taken longer. The first corpus consisted of 10 songs per decade from the decades 2010s to 1890s. The list for the top 10 list for each decade was taken from "DavesMusicDatabase.com". Whether these songs were really the best 10 songs of each decade wasn't all that important. The most important detail was that they were songs that were quite popular post release as this lets us draw a relation between popularity and grammatical patterns. It is worth noting that the second dataset which was the hip hop dataset, was a set of 24 songs chosen to add a unique perspective to the investigation. The songs were chosen in pairs where I attempted to use songs released further apart from each other where possible. There were a total of 12 artists chosen and 2 songs were chosen for each of them. The lyrics were collected in the same fashion as the first dataset. The artists were chosen based on who fit the following criteria. The artist had to be either an influential African American Artist or an influential hip hop artist. They also had to have an appropriate understanding of the African American perspective of life in America. Now in the case of Eminem, he was chosen because of who he grew up with and how this lead him to become a hip hop artist. He is also a cornerstone of hip hop and one of the greatest hip hop artists of all time. To snub him from the list due to his ethnicity seemed inappropriate.

4.2 Processing data

In order to use the NLP techniques that would allow us to extract meaning from the lyrics it is important to pre-process the data so that characters and/or words that don't contribute any meaning do not potentially confuse or interfere with the algorithm. In Figure 2 there are a number of different defined functions. Firstly the "remove_newlines" function removes any newline characters in the text and replaces it with a space between the last character of the previous line and the following line. This was done because the platform that had these lyrics stored had them

Figure. 2: Showing a list of defined functions

stored in centred paragraphs which lead to newline characters. There were also some weird escape characters similar to the newline character "\n" and so the function "remove_escape_characters" removes these characters from the corpus. Those two functions are the only ones used for pre_processing.

5 Results and Discussion

5.1 Sentiment Analysis

The results of the sentiment analysis showed a pattern of general positivity in the lyrics of the most popular song of each decade. This could be because many popular mainstream songs tend to

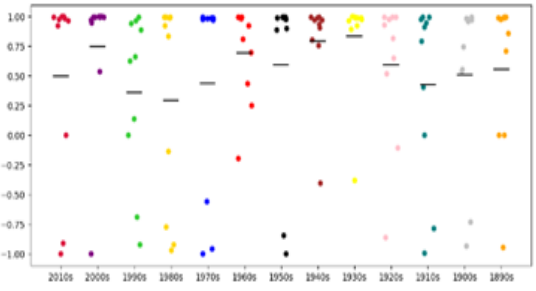

Figure 3: showing distribution of sentiments for the decade songs

deal with happier topics such as love or hope than in hip hop which tends to have songs about reflection and tells stories about struggle. For example Nas's "N.Y. State of Mind is a song describing the city of New York what was in his head when he lived in there when he wrote the song. In the 90s it was a city riddled with high violent crime rates where the streets where seeing the effects of the "Crack epidemic" . Mainstream music misses out on this side of the American story. Where rap as an art form often tells the story of individuals and how they got to where they are, mainstream often is more about experiences that everyone can relate to. Stories about love, partying or songs that are just plain catchy often dominate these platforms. There are some values missing due to weird interactions in the code and so there is a possibility that the extent to which some decades may have had negative values may have been inaccurate. However this is not really an issue since the decade most affected by this (1890s) is only missing 2 values in an era that wasn't of extreme historical importance when compared to some of the other decades. The decade with the least negative sentiment in American music was the 60s which would make sense as that was the decade of the hippies and the peace movement it would then make sense that out of all the eras that are supposed to have a general positive outlook on life, where everyone would be producing some of the happiest music would be the 60s. The hip hop sentiments were evenly spread. However there is a possibility that the classifier was biased towards the profanity expressed in the Hip Hop lyrics. The song **** in Paris had a value of -0.9993 due to the presence of a large amounts of profane language. For this reason the results for this section as it pertains to the second dataset have been omitted but can be found in the README.txt file.

5.2 Topic modelling

	Song 1	Song 2	Song 3	Song 4	Song 5	Song 6	Song 7	Song 8	Song 9	Song 10
2010s	0.9248	0.9994	0.9981	0.9997	0.9865	0.9776	-0.9991	-0.908	0.9635	0
2000s	0.9992	0.9941	0.9994	0.9981	0.5373	0.9966	0.9704	-0.9972	0.9478	0.9988
1990s	-0.6854	0.9405	-0.921	0.9971	0.8885	0.6249	0.6602	0.9661	0.1391	0
1980s	0.9953	-0.7731	0.9978	-0.1333	0.9913	0.9959	0.923	-0.922	-0.9698	0.8359
1970s	0.9989	-0.5538	0.988	0.9803	-0.998	0.9747	-0.956	0.9844	0.987	0.9893
1960s	-0.1943	0.987	0.9966	0.9985	0.8126	0.9906	0.4356	0.25	0.926	0.7003
1950s	-0.9954	0.9952	0.9959	-0.8407	0.8915	0.9947	0.901	0.9898	0.986	0.9778
1940s	0.9959	0.9392	0.9705	0.9062	0.8045	0.9697	-0.4019	0.9894	0.7555	0.9957
1930s	0.9274	0.9994	0.9897	-0.3792	0.9957	0.9878	0.9663	0.9885	0.8957	0.9795
1920s	0.9982	0.9962	-0.1033	0.8153	0.6537	-0.858	0.9975	0.9765	0.9279	0.5207
1910s	0.9937	0.9515	0.9777	0.9136	0	-0.9879	0.4082	0.7964	-0.7815	0.9944
1900s	0.9976	0.9623	-0.9303	0.5588	0.5373	0.9872	0.7469	0.9712	0.9865	-0.7311
1890s	0.8591	0.9918	0.9875	0.9972	0.9802	-0.9411	0.9962	0.7096	0	0

Figure7 : Sentiment analysis table for decade songs

Topic modelling is a technique that identifies the topics present in the latent space of the corpus of text as previously described and the probabilities of words associated with those topics. Now, depending on the number of topics you want the

	Probabilit	Probabilit	Probabilit	Probabilit	Probabilit	Probabilit	Probabilit	Probability
Word								
i	0.058	0.029	0.038	0.064	0.041	0.029	0.029	0.083
and	0.018	0.007	0.005		0.023	0.014	0.007	0.014
im	0.018	0.007	0.012	0.018	0.017	0.017	0.023	0.015
got	0.014			0.012	0.008	0.011	0.009	0.017
like	0.012		0.007	0.008	0.013	0.018		0.016
know	0.012				0.016	0.008		0.009
you	0.011				0.01			
dont	0.007	0.007		0.009	0.009	0.009		0.009
ni**a	0.007				0.009	0.008		

Figure 4: Showing the probability distribution of words in the hip hop corpus

machine to attempt to model, you can pass the function a different number. However, there is a limit to the number of latent topics identified. Therefore even if you give it an argument of n_topics = 200, it will only model 200 if there are 200 topics present in the latent space. Otherwise it will pass back as many as it can find. For the purpose of this project I chose 20 topics for both bodies of text. I chose this because the was what I discovered to be the limit for the number of topics in the latent spaces. For the hip hop topic modelling, the words that had the highest probability can be seen in Figure 4. With subject "i", occurring with the highest probability. Since rap is an art form that tends to describe the individual's perception and experience this is to be expected. In the list of popular songs by decade, the words the distribution of occurring words was spread slightly more. However, "I" still showed the highest probability of occurring.

What the output of the topic modelling looked like. The one thing both datasets demonstrated was that "I" occurred at the highest rate between both datasets. This is likely because music most often is used to describe one's experience. It is usually used to tell stories from the individuals perspective making the story personal.

5.3 POS Tagging

2010s	'PRP': 163, 'VBP': 237, 'JJ': 283, 'NNP': 474, 'NN': 611
2000s	'PRP': 231, 'VBP': 256, 'JJ': 372, 'NNP': 537, 'NN': 765
1990s	'JJ': 79, 'VBP': 110, 'PRP': 146, 'NNP': 162, 'NN': 218
1980s	'VBP': 98, 'PRP': 122, 'JJ': 156, 'NNP': 341, 'NN': 463
1970s	'PRP': 110, 'VBP': 113, 'JJ': 210, 'NNP': 334, 'NN': 472
1960s	'VBP': 124, 'PRP': 125, 'JJ': 162, 'NNP': 210, 'NN': 332
1950s	'PRP': 82, 'VBP': 87, 'JJ': 97, 'NNP': 307, 'NN': 337
1940s	PRP': 39, 'NNS': 52, 'VBP': 57, 'JJ': 76, 'NN': 176
1930s	'PRP': 79, 'VBP': 86, 'JJ': 101, 'NNP': 164, 'NN': 336
1920s	'VBP': 75, 'PRP': 84, 'JJ': 99, 'NNP': 166, 'NN': 184
1910s	JJ': 54, 'PRP': 54, 'VBP': 83, 'NNP': 178, 'NN': 244
1900s	'PRP': 62, 'VBP': 65, 'JJ': 157, 'NN': 249, 'NNP': 317
1890s	'PRP': 95, 'VBP': 107, 'JJ': 165, 'NNP': 263, 'NN': 306

Figure 5 showing decade songs POS distribution

POS tagging revealed that the general distribution in terms of types of words used in mainstream popular has relatively stayed the same. No matter what era Singular nouns were always the most used POS. After that though, the distributions varied. Generally, proper nouns were the second most used with adjectives coming in at the third most used part of speech. The third and fourth most used POS saw the most variation. This was mostly because in these two placements, categories were used at very close levels to each other. Often with 20 instances being the difference between one POS being used more often than another. The distribution indicated a tendency for popular music to focus of people, objects events and places. This is to be expected since more often than not the subject is just as important if not more so than the description of said subject. This was essentially the same in the HipHop corpus. The distribution was essentially exactly the same as shown in figures 5 and 6.

6 Limitations and Future Work

Although it was possible to draw a number of conclusion from the results published in this paper there is still room for improvement. Creating a larger corpus of data and possibly increasing the number of songs in the African American/hip hop/ soul genre would allow for better and more accurate observations to made from the dataset. It is certainly possible due to the size of the dataset certain types of biases may have been made more apparent due to selection bias (since artists that should have been included may have been left out).

Furthermore, the mainstream and African

L.Hill	'RB': 39, 'VBP': 43, 'JJ': 53, 'NNP': 107, 'NN': 113
Nas	VBP': 77, 'JJ': 106, 'NNS': 108, 'NNP': 166, 'NN': 253
Tupac	'PRP': 61, 'VBP': 71, 'JJ': 88, 'NNP': 116, 'NN': 194
Eminem	'VBP': 122, 'JJ': 124, 'PRP': 128, 'NNP': 156, 'NN': 274
Biggie	'PRP': 73, 'VBP': 96, 'JJ': 121, 'NN': 166, 'NNP': 221
Wu-Tang	'NNS': 71, 'VBP': 79, 'JJ': 119, 'NNP': 229, 'NN': 304
Kendrick	'PRP': 98, 'VBP': 98, 'JJ': 104, 'NNP': 117, 'NN': 234
Jay-Z	'NNS': 65, 'VBP': 81, 'JJ': 96, 'NNP': 148, 'NN': 164
Snoop	'PRP': 53, 'NNS': 65, 'JJ': 118, 'NN': 170, 'NNP': 179
Ye	'NNS': 45, 'PRP': 54, 'JJ': 74, 'NNP': 102, 'NN': 204
Drake	'PRP': 54, 'VBP': 62, 'JJ': 75, 'NNP': 112, 'NN': 175
J-Cole	'VBP': 68, 'JJ': 82, 'PRP': 83, 'NNP': 103, 'NN': 221

Figure 6: Showing the POS distributions for different artists

American experiences are not the only sides to the story as America has more ethnicities and subcultures that make up it's music culture. Increasing the number of artists and individuals as well as increasing the diversity of these individuals may have allowed for a more complete observation to be made regarding the changes in language in popular American music. Furthermore, if possible manually assigned scores of sentiment could have been assigned to the HipHop songs or some score adjustment could have been applied in order to allow for a comparison of this category to the other popular works. Lastly a wider range of HipHop artists could have been chosen.

7 Conclusion

Based on what was observed in this project it is appropriate to say that music as a whole on the grammatical level has not changed as much as one would think. It was an art form that was used to tell a story of a subject or number of subjects and seems to have remained that way. Furthermore, HipHop and mainstream music have shown great similarities in most categories. In the future, maybe when there is time to find a way to adjust for the presence of specific words due to the nature of HipHop, it would be interesting to compare the sentiments of the popular HipHop songs to the sentiments of the popular mainstream songs. As for the overall positivity or negativity of popular works, generally, mainstream music has more

positive music that goes popular than it does have negative. Probably because popular music in the mainstream tends to be catchy and upbeat. This is left up to speculation for now. In this paper we have shown the differences and similarities in different musical works from the 1980s to 2019 across a number of different genres.

9 References

[1] Top 10 songs by decade: https://davesmusicdatabase.blogspot.com/2012/04/top-songs-by-decade-1900-present.html

[2] Most influential HipHop artists: https://www.standard.co.uk/culture/music/the-most-influential-hip-hop-artists-of-all-time-a3863356.html

[3] American history: American History Timeline 1800-1900 (american-history.net)

[4] African American history: Black History: Facts and People | HISTORY.com - HISTORY

[5] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh, Natural language processing: State of the art, current trends and challenges, Natural language processing: state of the art, current trends and challenges (springer.com)

[6] Code for plotting a vertical scatter: https://stackoverflow.com/questions/44982574/how-to-plot-vertical-scatter-using-only-matplotlib