Topic Modeling with Song Lyrics

by Alwine Balfanz (3728215), Julia Güttler (3727667) and Lisa Wagner (3740533)

Abstract

Music comes in different types and styles and hence can be categorized into different genres based on certain criteria. This project focuses on genre specific lyrics and the topics addressed. We use LDA Topic Modeling to answer the question, whether we can tell music genres apart based on their lyrics. In order to examine the differences in the topics distribution we built our own lyrics corpus with songs from a total of 14 genres. After calculating several topic models we set the number of topics to 12. We visualized the topic proportions of each genre with regard to comparing the genres with each other. It can be observed that the genres vary in their topic distributions. Some genres are dominated by one or two topics, others have more evenly topic proportions.

1. Introduction

Topic Modeling is a common practice in the field of Natural Language Processing (NLP) and text mining. It is usually used in text analysis to determine common themes. Topic models are statistical models, which are used to discover abstract topics. Although, coming from the world of text analysis in a more classical context, topic models had successful implementations in other fields, such as medical document mining.

In the case of musical information retrieval, the focus on lyrics gained attraction in recent years. There have been different analyses on the textual level from lyrics, mostly concerning the underlying syntactic or semantic structure of lyrics, semantic value analysis or other common text mining methods like in Werner (2012). Topic modeling on lyrics is a fairly new method that has been mostly concentrating on a few music genres or just pop music in general. Therefore, we want to focus on finding out which topics are prominent in 14 different music genres and whether the distribution of those topics over the genres can tell us something about the genres' backgrounds. We intend to answer those two research questions:
Q1: Can we tell music genres apart by identifying typical topics based on their lyrics?
Q2: Is it possible to trace back the occurrence of certain topics to the origins of those genres?

In this project we analyzed the following genres: Pop, Hip-Hop, Rap, RnB, Soul, Rock, Country, Folk, Blues, Jazz, Heavy Metal, Reggae, Gospel, Spiritual music and children's songs. Due to the fact that children's songs did not provide enough material, we included Disney songs (from Pixar and Disney animation studios movies) into that category.

We built our own lyrics corpus for this project, with lyrics from Genius[1]. The cleaned data set of songs has about 3500 songs (from all genres). The topic modeling was done in R with the Latent Dirichlet Allocation (LDA) algorithm.

[1] https://genius.com/

2. Related Work

Although Music Information Retrieval has gained traction in the digital humanities in the last 20 years, researchers' focus has concentrated predominantly on the audio components of songs and pieces. Lyrics topic modeling especially has only rarely been done in the past. In existing research regarding topics in song lyrics there has been song classification into predetermined or manually annotated subjects or extraction of topics from the whole lyrics corpus without a matching of the topics to the music genres.

Kleedorfer et al. (2008) created a music indexing algorithm using non-negative matrix factorization to identify topic clusters after transforming the lyrics into a vector space model.

Sasaki et al. (2014) implemented a lyrics retrieval system called LyricsRadar for interactive browsing to find songs with similar topics using latent dirichlet allocation. They developed a graphical user interface to help visualizing lyrics on two levels; a topic radar showing the meaning of lyrics and a two-dimensional plane locating songs with similar lyrics.

Choi et al. (2016) classify songs into eight predetermined subject categories by tf x idf features extracted from the songs' lyrics and user interpretations of the lyrics. They use two SVMs, a kNN and Naive Bayes classifier with 10-fold cross validation and reach over 70% accuracy on some categories noting that the classification based on the user generated interpretations is more accurate than that of the lyrics themselves. This research is slightly altered in Choi and Downie (2018) who transform the words in the lyrics and user interpretations into vector representations and use those as input for a Naive Bayes Classifier to classify into the same predetermined topics.

Topic modeling on Indonesian songs has been done by Laoh et al. (2018). They chose songs in the complex and poetic language Bahasa and extracted 10 topics. The documents are represented in a document-term-matrix which is the input for the LDA.

Denzler (2021) performs LDA topic modeling on 120,000 songs representing their lyrics in a bag-of-words format and receiving 6 topics as a result.

Antoniou (2018) performed topic modeling with LDA on a 380,000 song data set with genre information for each song with the result of only receiving two topics. Almost all hip-hop songs were assigned one topic whereas almost all songs from the other 10 genres were assigned the second topic.

In this project we want to further explore the matching of topics in the lyrics to the songs' genres to see whether there are prominent topics in the lyrics of certain genres and whether a genre assignment based on the appearing topics in a song could be possible.

3. Methods

The methods that have been used in this project are methods of text mining, NLP and statistical modeling. The data collection and a part of the data cleaning has been done in the programming language python. The other part of data cleaning and preprocessing, as well as the topic modeling (LDA) and visualization have been done in R.

3.1 Data Collection

To build a corpus of song lyrics the top albums for each chosen genre as ranked by LastFM were used. HTML files of the LastFM websites were scraped with the program wget used in the terminal with the lines

```
wget https://www.last.fm/tag/GENRE/albums
```

and

```
wget https://www.last.fm/tag/GENRE/albums?page=PAGE_NUMBER.
```

The names of the albums and artists were scraped from these HTML files using the python library beautifulsoup4 (script compileAlbumLists.py). The artist and album names were saved in CSV files to be loaded in the lyrics scraping program.

The lyrics are gathered with the Genius API and corresponding python library lyricsgenius, which provides a function to retrieve lyrics of a whole album when given the album and artist name (script getAlbumLyrics.py). Those were loaded from the CSV files and subsequently the lyrics were scraped and saved in a data frame. Additional metadata containing the song name, and date of release were saved in the same data frame.

One problem we encountered was that there were only very few children's songs on LastFM that did not throw an error at the attempt of retrieving their lyrics from genius. Often the artist name listed on LastFM for soundtrack albums of Disney movies for example was "Various Artists" which is why it was not found on Genius. Subsequently, we had to search for these albums on Genius manually and correct the name of the artist in the lists. Since the search for these albums mainly from Disney and Pixar movies happened manually and it was clear therefore that those songs were children's songs, the tagging script was modified to tag a song with the "kids" tag, even if this was not in or on top of the list of tags.

As it was already apparent that some albums appeared in the lists of several genres, and songs can generally be quite different even on the same album, the next step was to assign a tag to each song individually (script tagSongs.py). For this step LastFM was used again, this time the LastFM API and python library pylast. On LastFM users can assign tags to songs. Those can be genre tags like "pop", "reggae", or "gospel", but also individual tags which do not represent a genre or class the song belongs to. Each tag has a specific weight according to how many times the song was assigned the specific tag. A pylast function returns a list of the top tags with their respective weights when given the name of the song and the artist. This list was extracted for every song. The song was tagged with the highest ranked tag which was in our list of accepted genres. In the case that the list of top tags did not contain one of the accepted tags, the song was assigned the tag "noTag" to signal that the song did not belong in the corpus. Occasionally errors occurred when trying

to retrieve the list of top tags from LastFM. In that case the song got the tag “topTagsError”.

3.2 Data Cleaning and Preprocessing

The first step of the data cleaning was to clean the lyrics. It is pretty common that the lyrics from genius.com come with strings such as ‘[Chorus]’, ‘[Verse 1]’, ‘[Verse:]’ and so on. These had to be removed before anything else. For this a python script (preprocessing.py) looked for lines with a ‘[’ at the beginning, as the above mentioned strings were mostly in brackets. If this was the case, the line was removed from the lyrics. Other variations were removed manually. There were also strings containing ‘Embed’ with random numbers before that string. These had to be removed manually as well as other variations of the song part indicator words like ‘(Chorus)’, ‘(Verse 1)’, ‘Verse:’. The corpus was saved as tagged_corpus_preprocessed.

After that other preprocessing steps, like omitting certain songs, lemmatization and stemming, had to be done in R. Therefore the corpus was read as a data frame. First, we removed all songs with the ‘noTag’ and ‘topTagsError’ in the tag column. Then we had to check whether the lyrics were in English. This was done with the textcat package of R. This package is not 100% correct in detecting the language of the lyrics. It performed better, when the other languages were added to the language profiles. But even with that step, the detection was not perfect.

After that, we built a corpus object with the lyrics and song IDs. The lyrics were tokenized and punctuation, digits, and symbols were removed. We also put the words to lowercase and lemmatized them. This was done with the ‘baseform_en.tsv’ file. We also removed common English stopwords and also tokens like ‘la’, ‘oh’, ‘yeah’, since they are frequently used in songs, but add no value to meaning and topics.
Collocations were also calculated and added to the corpus tokens.
Then we sampled 100 songs from each genre (we added Spiritual music to Gospel, since they are similar and Spiritual music only had 6 songs in the corpus) and stored

the indices in vectors. After that we built a DTM object and removed objects with 0 as value.

3.3 Topic Modeling

For the Topic Modeling we used the topicmodels package from R and the LDA algorithm. After various tested models we decided that the best results were calculated with K set at 12 topics, an alpha value of 0.2 and the number of iterations at 500 (topicModeling.R). This decision was made with the look at the top 10 words per topic. We also tried to make a better assumption about the parameters using the LDAtuning package, but we did not get a local minimum. The first local maximum from Deveaud2014 was at 12 topics, which gave us a clue for the best number of topics.

3.4 Genre Backtrace

In order to examine the second research question in more detail we took a look at each single song from the sampled data set and compared the calculated theta values with the theta values of each genre which was the average of all theta values of the songs belonging to this genre. The corresponding source code can be found in genre_backtrace.R. As a measure of comparability we chose the cosine similarity. The function cos_sim takes as parameter the song id and then calculates the cosine similarities with each genre based on their theta values. The similarities are sorted in descending order. The top 3 most similar genres are returned. This function is applied to all songs of the sample data set. Therefore we store the top 3 most similar genres and the correct genre in a data frame. In a next step it is checked whether the correct genre is included in the top 3 genres. Finally, the share of the songs with correctly identified genre is calculated.

3.5 Visualization

First we visualized wordclouds of the topics, to get a better idea of the overall themes in each topic. This was done with the wordcloud2 package. For that two

variables were initiated: topic and word. Topic is for the number of one topic and the word variable is for the most used word within that specific topic. With those variables the selection of the topic can be overwritten and the creation of wordclouds of other topics is fast and easy. (All wordclouds can be found in the word clouds folder of the Git Repository.)

To get a first idea of how the topics proportions are in the different genres, we calculated the overall proportions of the topics across all genres, additionally we calculated the number of songs where the topics were most probable.

After that we visualized the topic proportion of every genre separately. We extracted the theta values corresponding to the song IDs of the genres and made new data frames for each genre. Then we calculated the mean value of every topic within the genre. This data was visualized. Lastly, we visualized the proportions of topics for all genres, to compare them better. (All graphics can be found in the Topics Proportions folder of the Git Repository.)

4. Results and Discussion

In this section we will present our results and interpret and discuss them with the help of the diagrams that visualize topics and topic distributions and evaluate the research questions.
Before we go into a detailed analysis of the topic distribution over the genres, here are some key observations of our experiments.

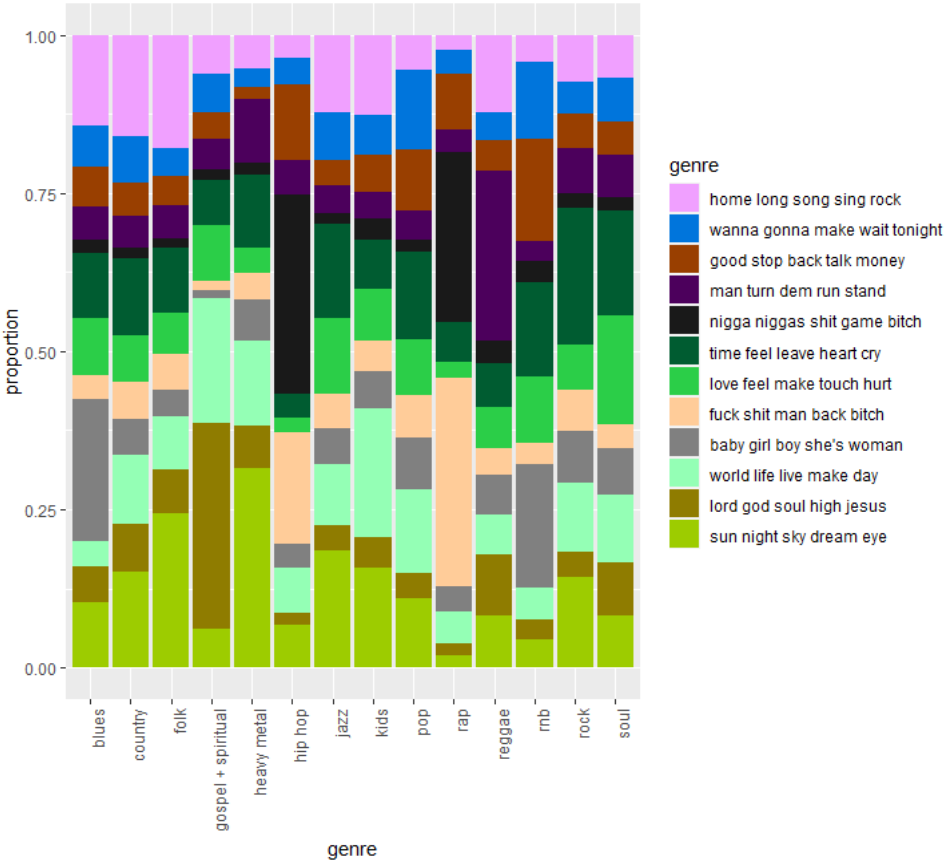

Figure 1: Bar plot of topic proportions over all genres.

There are certain topics that appear steadily in most of the genres. Unsurprisingly, they contain the subjects of love and heartbreak, but also nature and life. Other topics restrict themselves to a few genres only, like religion and belief or obscenities and profanity. A diagram showing the distributions of all topics over all genres may be seen in figure 1.

Additionally, there are genres in which the topics are more evenly distributed than in others like pop (see figure 2), country, jazz, soul, or R'n'B, whereas in other genres like hip-hop (see figure 3), rap, and reggae, but also gospel, and heavy metal, there are only one or two strong topics and the others are negligible.

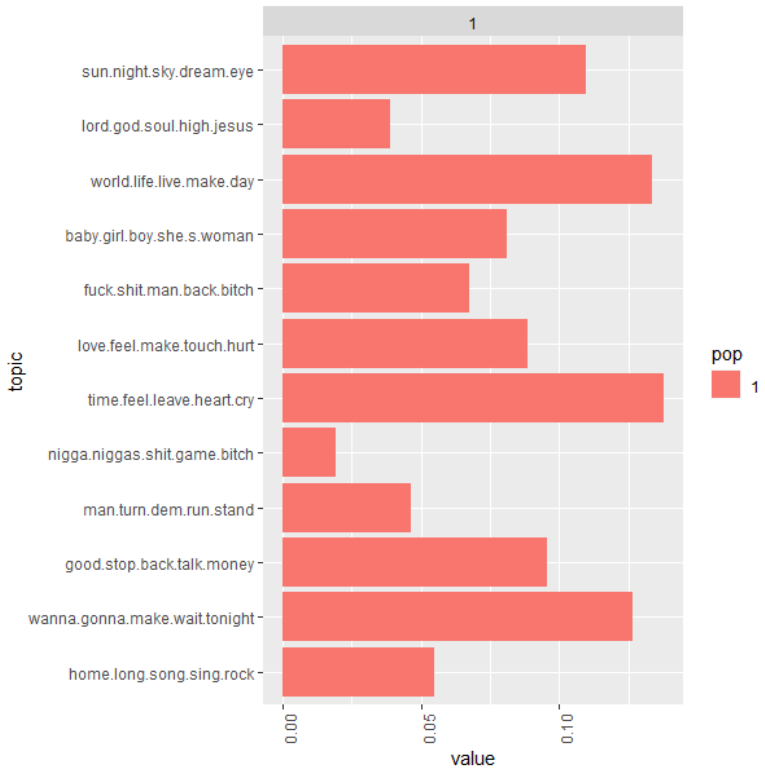

Figure 2: bar plot of topic distribution of pop songs.

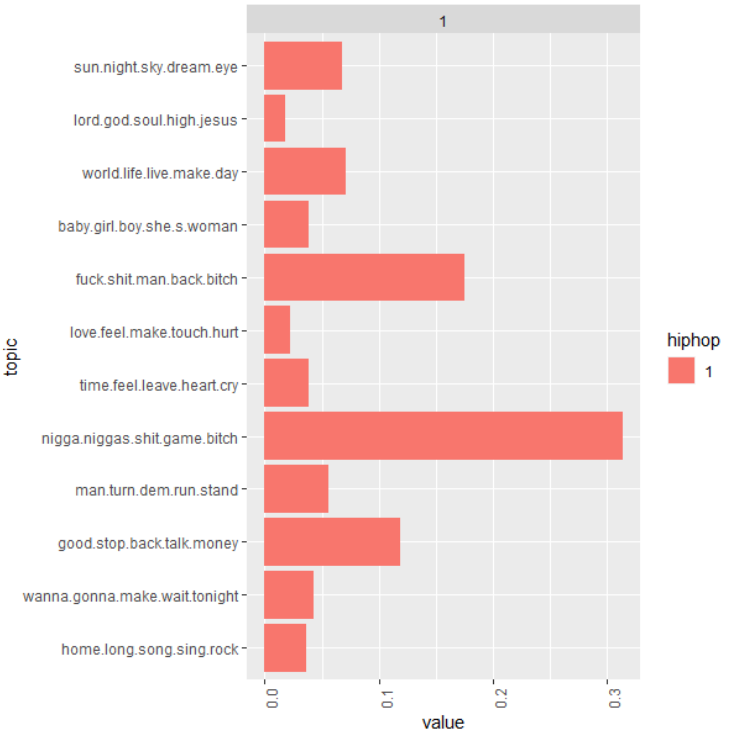

Figure 3: Bar plot of topic distribution of Hip-Hop songs.

4.1 Topic Analysis

The first topic can be described with the word “home”. Its distribution over the different genres is fairly steady but it appears most in folk and country songs with around 17% of words in those genres’ lyrics being associated with the topic. As folk is a very widespread genre which includes traditional as well as contemporary folk it is hard to draw conclusions about the genre as a whole, but the importance of (national) identity and culture seems to be present in most folk songs. The lyrics of country music, which had its roots in American folk as well as blues and spiritual music, often concerns the American way of life, living in a small town or on a farm, and also patriotism and being proud of the home country. This might be the reason why the topic “home” is prominent in country music.

The topic also appears in kids’ songs, around 10% of words in kids’ songs’ lyrics are associated with it, presumably because kids can relate to it and it is easy to understand. It could also relate to “The Hero’s Journey”, a popular trope in movies, in which the main character leaves home to go on an adventure and eventually returns home. This storyline could be a source for songs about leaving, valuing and returning home. The topic appears least in gospel, heavy metal, hip-hop, rap, pop, and R’n’B songs.

The words in topic two are not very substantive. Some of the most popular words like “gonna”, “gotta”, “make” do not carry much meaning regarding a specific topic and should presumably be counted as stop words (for instance “gonna”, the short form for “going to” which are both stop words). The words could describe something like “party” with indicating words like “tonight”, “dance”, and “play”. Moreover, the topic is not very important in the genres, mostly so in pop and R’n’B songs with around 12% of words relating to it. Those could just be dance or party songs, where the lyrics are less important than the sound of the music and the ability to dance to it. In all other genres less than 8% of the words are connected to topic two.

In the words for topic three there is no real topic conceivable. The most prominent word by far is “good”, which does not really say much, and the other words are not very insightful either. Yet there are differences in the topic distribution. It appears

most in R'n'B, hip-hop, rap, and pop music, and least in metal, jazz, spiritual, folk and reggae.

Some less prominent words ("child", "friend", "son", "daddy", "brother", "mother") in the wordcloud of topic four (see figure 4) suggest that it could be about family and friends, however the most important word is "man", which seems to be a form of address. Other words like "dem", "nuh", "fi", and "mi" sound like a Jamaican accent and the topic more than 25% of words in reggae songs are associated with the topic and not more than 10% of words in all other genres. This topic seems to have picked up the linguistic difference of the Jamaican accent, which is most used in the lyrics of reggae songs.

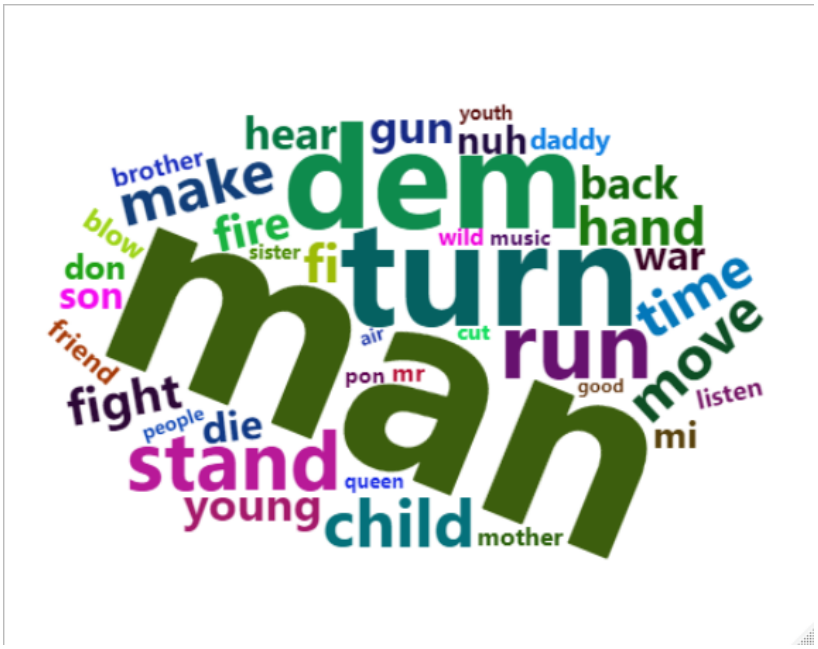

Figure 4: Wordcloud of topic 4.

Topic number 5 also seems to have detected linguistic differences of two genres rather than a real subject. It contains (mostly race and gender based) obscenities, swear words and derogatory language as well as words indicating violence. The topic is by far the most prominent one in hip-hop (more than 30% of words), and rap (more than 25% of words), and almost nonexistent in the other genres. This is more or less what we had expected. Rap and hip-hop artists often come from low-income high-crime backgrounds where "street language" and being vulgar might be more

common. They often want to be sincere and not censor themselves or mince their words.

Topic six is about heartbreak and lovesickness. It is distributed fairly steadily over the genres though least present in hip-hop, rap, and gospel songs. It appears most in rock (more than 20%), pop (around 17%) and soul (around 16%). We had anticipated that this topic, which is very popular, would appear in every genre. Writing songs could be a coping mechanism for artists to deal with their lovesickness. Perhaps it makes sense that the topic is less prominent in gospel and spiritual music because its lyrics are less concerned with interpersonal relationships.

The seventh topic is about love more generally. “Love” is by far the most important word in the word cloud. Like the previous one, this topic appears in all genres, most in soul songs (around 17% of words) and least in hip-hop, rap, and metal songs. As stated before, gospel and spiritual music might be less about interpersonal relationships, but the uses of “love” in those lyrics could concern the love of or towards Jesus or God.

Topic number 8 is, as topics 4 and 5, not so much of a “topic” per se, but more an indicator of the language used, which again contains many swear words, although they may be considered more “benign” and less derogatory than the words in topic 5. Unsurprisingly, this topic is again conspicuous in rap (its most important topic with more than 30% of words relating to it) and hip-hop (around 17%). Not more than 8% of words in the other genres’ lyrics are connected to the topic, and it is least prominent in spiritual music.

“Baby” is the most prominent word in topic 9, which makes it probable that the songs directly address the person it is about. The topic could be called “man and woman” or “relationships” in general, as words like “girl”, “boy”, “woman”, and “love” are important too. It is the most popular topic in R’n’B (around 19% of words) and blues lyrics (25%). Blues originated in spiritual and folk music and while in the beginning the lyrics were often about economic difficulties, it later became more about relationships and sexual issues. R’n’B emerged from blues, soul and gospel music with influences from rock and roll. Given that, it is not surprising that this topic is

most important in blues and R'n'B, while it appears considerably less in the other genres, least so in spiritual music, hip-hop, rap and folk.

The tenth topic seems to be about "life". While "world" is the most important word in the word cloud, many others also have a positive connotation such as "life/live", "free", "hope", "dream", "good", "change", "grace". The topic is most prominent in kids songs (more than 20%) and spiritual music (around 20%). We had anticipated that children's songs would have more positive connotations and might be about adventures and the joy of living. The prominence of the topic in spiritual music suggests that those lyrics are about life and all the beautiful things in the world which were given by God. This topic is also fairly prominent in most other genres except blues, which often concerns economic and personal woes, and R'n'B, which had origins in blues, and rap music.

The eleventh topic is very straightforward. It concerns religion and belief, with "lord", "god", and "jesus" being the most important words, and unsurprisingly appears most in gospel and spiritual music (around 32% of words). More than in other genres it also appears in reggae and soul lyrics as reggae is connected to the Rastafari religion and soul had roots in gospel, blues and jazz music. However, it is slightly surprising that the topic is not more popular in country music, where only around 8% of words are associated with it, because as stated before, country has roots in spiritual music and there are often connections in country songs to the personal belief of the artist. The topic is almost nonexistent in hip-hop and rap lyrics.

Finally, topic number 12 is about nature, although the words in the word cloud (see figure 5) indicate that the connotations of this topic can be very different. Words like "sun", "sky", "blue", "shine" have a more positive connotation in contrast to words like "night", "die", "dark", "fall", "end", "burn". That might be the reason why heavy metal songs have the highest share of this topic (more than 30%). Since metal is often associated with dark and depressing topics, it is possible that those lyrics rather make use of the words with negative connotations. Folk songs are a lot about nature as well (around 24%). Although as stated previously it is hard to draw conclusions about the genre as a whole because of how widespread it is, it is conceivable that nature can play a big part in the culture and identity of a community. The topic is also prominent in country, jazz, kids, and rock music (around 15% each). While country

and kids songs are likely to sing about the positive connotations of nature, rock lyrics might be closer to those of metal songs.

Figure 5: Wordcloud of topic 12.

4.2 Research Questions

Q1: Can we tell music genres apart by identifying typical topics based on their lyrics?

Certain topics appear overwhelmingly in one of the genres (religion in gospel, obscenities in hip-hop rap, Jamaican accent in reggae) so there is a high chance that a song whose lyrics cover one of those topics belong to the certain genre. For some topics it would be possible to exclude genres for which the topic has a low probability. Hip-hop and rap lyrics are unlikely to cover the topics religion, home and love, gospel music is unlikely to cover relationship issues, or profanities.

With this number of genres (14) and topics (12) however, it would not be possible to assign a genre to a song purely because of a topic appearing in its lyrics, as the topic distribution chart also shows.

If we had a topic distribution over several songs that all belong to the same genre, it might be possible to figure out which genre we are dealing with on the basis of the distribution of the different topics.

Q2: Is it possible to trace back the occurrence of certain topics to the origins of those genres?
The topics of religion and belief as well as the positive view on life in gospel and spiritual songs seems logical.

The prominence of swear words and obscenities in rap and hip-hop music is not surprising given the previously described background of these genres.

Kids songs should be about subjects children can understand such as home, nature and the beauty of life and our experiments show they often are.

The appearance of religious topics in soul music can be traced back to the genre's roots which are partly in spiritual music, and reggae music is connected to the Rastafari religion, which can explain the share of religious lyrics in reggae songs.

That R'n'B partly emerged from blues can be seen in the shared topic of relationships, which appears in those genres the most.
In short, in some cases the occurrence of a topic in a genre makes perfect sense because it can be traced back to the roots and origins of the genre.

The results for the genre backtrace of songs was statistically not outstanding. In our case for about 55 % of the songs the correct genre occured in the top 3 most similar genres. This means that the topic distribution of a single song is insufficient to determine a song's genre. It rather gives only a hint in which direction it is going. More characteristic genres like gospel and hip-hop might be easier to identify since they have very few dominating topics.

5. Conclusion

The purpose of this project was to identify genre specific topics in song lyrics using LDA Topic Modeling. Therefore we built our own lyrics corpus. We focused on 14 genres and subsequently gathered the lyrics with the Genius API. After Cleaning and preprocessing the data several topic models were calculated, finding that for our objective a number of 12 topics, an alpha value of 0.2 and 500 iterations had the

best outcome. To visualize our findings we made use of wordclouds and bar plots showing the most common words of each topic and the proportions of topics in each genre. The graphics show that the genres differ in their topic proportions. Some of the genres are characterized by one or two rather dominant topics like the religion and belief topic in gospel and spiritual lyrics or the topic 5 containing obscenities and swear words which is prominent in hip-hop and rap songs. Other genres have a more even distribution of topics for instance pop, country or jazz. A detailed topic distribution over a document could allow us to assign a genre to the document with some confidence. In further work this classification problem could be studied in more detail. In the decision making process some other metrics could be taken into consideration in order to increase accuracy. Vocabulary and stylometry may be suitable in this case.

References

Antoniou, M. (2018, June 27). Text analytics & topic modelling on music genres song lyrics. *Towards Data Science*. https://towardsdatascience.com/text-analytics-topic-modelling-on-music-genres-song-lyrics-deb82c86caa2

Choi, K., & Downie, J. S. (2018). Exploratory Investigation of Word Embedding in Song Lyric Topic Classification: Promising Preliminary Results. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 327–328. https://doi.org/10.1145/3197026.3203883

Choi, K., Lee, J. H., Hu, X., & Downie, J. S. (2016). Music Subject Classification Based on Lyrics and User Interpretations. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–10. https://doi.org/10.1002/pra2.2016.14505301041

Denzler, T. (2021, May 14). What's in a Song? LDA Topic Modeling of over 120,000 Lyrics. *Medium*. https://tim-denzler.medium.com/whats-in-a-song-using-lda-to-find-topics-in-over-120-000-songs-53785767b692

Kleedorfer, F., Knees, P., & Pohle, T. (2008). Oh Oh Oh Whoa! Towards Automatic Topic Detection in Song Lyrics. In J. P. Bello, E. Chew, & D. Turnbull (Eds.), *Proceedings of the 9th International Society for Music Information Retrieval Conference* (Vol. 9, pp. 287–292). Drexel University.

Laoh, E., Surjandari, I., & Febirautami, L. R. (2018). Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation. *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, 270–274. https://doi.org/10.1109/ICISCE.2018.00064

Sasaki, S., Yoshii, K., Nakano, T., Goto, M., & Morishima, S. (2014). LyricsRadar: A Lyrics Retrieval System Based On Latent Topics Of Lyrics. In H.-M. Wang, Y.-H. Yang, & J. H. Lee (Eds.), *Proceedings of the 15th International Society for Music*

Information Retrieval Conference (Vol. 15, pp. 585–590). Zenodo. https://doi.org/10.5281/ZENODO.1418075

Werner, V. (2012). Love is all around: A corpus-based study of pop lyrics. Corpora, 7(1), 19–50. https://doi.org/10.3366/cor.2012.0016