Social Data Science

Department of Economics

Exploring Song Lyrics

- a quantitative text analysis of the differences in language, topics and sentiment of lyrics across musical genres

Contributions

According to the curriculum we are required to define the primary writer of every part of our paper, however all students have contributed equally to every section in the paper. The abstract, introduction, motivation, literature review, section introductions and concluding remarks were written jointly.

Exam number 77: Section 4.1 - Line: 1-8, Section 4.3 - Line: 1-13, Section 5.1 - Line: 1-12, Section 5.2.1 - Line: 31-46, Section 5.2.2 - Line: 31-46, Section 6.1 - Line: 1-15, Section 6.2 - Line 14 - 26, Section 7 - Line 19-26

Exam number 92: Section 4.1 - Line: 9-15, Section 4.3 - Line: 14-26, Section 5.1 - Line: 21-21, Section 5.2.1 - Line: 47-58, Section 5.2.2 - Line: 47-50, Section 6.1 - Line: 16-30, Section 6.2 - Line 27 - 36, Section 7 - Line 27-33

Exam number 105: Section 4.2 - Line: 1-16, Section 4.4 - Line: 1-16, Section 5.2.1 - Line: 1-15, Section 5.2.2 - Line: 1-15, Section 5.2 - Line: 1-14, Section 6.1 - Line: 31-40, Section 7 - Line 1-8

Exam number 87: Section 4.2 - Line: 17-32, Section 4.4 - 17-32, Section 5.2.1 - Line: 16-30, Section 5.2.2 - Line: 15-30, Section 5.2 - Line: 15-28, Section 6.2 - Line: 1-13 , Section 7 - Line 9 - 18

ECTS points: 7.5
Keystrokes (including spaces): 41,869 (Charcounter)
Date of submission: 01/09/2018

Exploring Song Lyrics

- a quantitative text analysis of the differences in language, topics and sentiment of lyrics across musical genres

Department of Economics, University of Copenhagen

1. September 2018

Abstract

In this paper, we seek to analyse the differences in language, topics and sentiment of lyrics across musical genres. The four investigated genres are pop, country, rap and rock as they give a good representation of different music tastes. In order to perform the analysis we use web scraping as data collection tool, and scrape the Genius website (genius.com). In doing so, we identify several textual features that differ across the genres, as the word length and the share of unique words. We find that the differences between pop, country and rock can be blurry, while our analysis indicates very special characteristics for the rap genre. This separation between rap and the rest of the genres becomes clear in the word length, share of unique words, topics covered and the sentiment analysis which indicates that rap appears to be more negatively charged, than the other genres. Using a logistic regression model, we obtain the highest precision when predicting rap lyrics, while the prediction of the other genres is lower. By incorporating the features found to be important in our analysis we achieve a higher prediction and thus argue for including the features in addition to the bag-of-words.

Contents

1 Introduction

"Music expresses that which cannot be said and on which it is impossible to be silent."
- Victor Hugo

From medieval troubadours to contemporary mega DJ's, music has been a part of our everyday lives for centuries. From country to rap, different musical genres reflect cultural and socioeconomic backgrounds. Several studies, e.g. Marshall & Naumann (2018), have found that music preferences are associated with gender, race and socioeconomic characteristics of the individual. In addition, results from a range of studies have indicated that music can affect feelings, identity and behavior (American Academy of Pediatrics 2009). These findings suggests that researchers can study people by studying the music that they listen to.

In this paper we investigate 3,441 musical lyrics from genius.com. Specifically we compare lyrics from four different musical genres: rock, pop, rap and country, and seek to answer the question: what can quantitative text analysis reveal about the content of song lyrics across musical genres? Our objective is to show how these tools can be utilised to acquire knowledge about the lyrics - knowledge that potentially could be combined with user data in future studies and investigate *why* individuals with certain characteristics tend to prefer certain genres, and *why* music may affect feelings and behavior differently depending on the genre.

We analyse three different aspects of the lyrics; the language, the topics and the sentiment. To examine the variations in the language, we calculate the number of words and unique words and other measures for lyrics within each genre. Furthermore, we use word clouds to visualise the most frequently used words within the different genres. To get a deeper understanding of the different genres we estimate a topic model to see if the topics vary for the different genres. Lastly, we conduct a sentiment analysis of the different songs, to see if the sentiment differs between the genres. We end our analysis with a proof of concept, where we estimate a machine learning model to identify the genre of unknown lyrics based on our findings in the analysis. Our objective is not to build as precise a classifier as possible, but instead to examine what we can learn about different genres of music from the lyrics alone, and thereby disregarding sound and image.

In general our results show that there exist some textual differences between the genres. Especially we find that the rap genre varies from the other genres in many of the investigated features. We find that rap lyrics are in general longer, have a larger vocabulary, deals with different topics and use a more negatively charged language. We find that the remaining three genres rock, pop and country are more similar, particularly when it comes to the topics.

2 Motivation

Why is it interesting to investigate lyrics across different musical genres? Several studies within the fields of psychology and sociology have investigated music in relation to behavior, identity, social movements, health and much more. As mentioned in the introduction, studies have found that preferences for music vary with characteristics such as gender, race and socioeconomic characteristics of the individual and related research have suggested that music can have the ability to affect individuals in different ways.

A study from 2017 found that the average American listened to music 32 hours a week (Forbes Online, accessed 31-08-2018). This number was higher than ever before and serves to illustrate the increasing potential impact that music may have on our lives. However, the themes, language and musical expression may differ across genres. If this is the case, then analysing these different aspects of the genres, may give an indication of how different music genres might affect people.

For studies engaged in the effect of music on behavior, isolating the causality is a major obstacle. It may be that a person who listens to aggressive music with violent topics will be affected by the content and carry out violent actions. However, it may also just be that a person who is violent prefers to listen to this kind of music. In the latter case there is a correlation between violent behavior and violent actions, but music may have had no causal effect at all. This paper is not concerned with isolating a causal connection between musical genres and behavior, but rather answering the preliminary question of whether there is a difference in lyrics across different musical genres. However, the aspects we have chosen to analyse are all factors that one could imagine holding potential for affecting individuals, if a causal relation does exists. For instance one could imagine that lyrics containing e.g. harsh language and negatively charged words and sentences, would have a different effect on listeners, than music with more positive sentiments. Likewise, music with a large vocabulary and many characters per word might have a more positive effect on the language development of children and youth, than music with very simple and repetitive language.

Web scraping enables us to obtain large amounts of data from the internet. By applying web scraping techniques to websites with song lyrics, it is possible to quickly collect large amounts of song lyrics, that in terms of its size can give a broad picture across the different genres. Additionally, natural language processing enable people to work with and quantify very large amounts of text. These new and groundbreaking methods can be utilized to investigate large amounts of lyrics and examine if there exist differences across the genres.

In this relation, a further motivation for the project is that the approaches considered in this paper are relevant also in other fields than music. The approaches can also be used on for example Twitter and Facebook posts. Generally quantitative text analysis is developing rapidly and hold huge potentials for future research in many areas.

3 Brief literature review

Many earlier studies have treated scraping and classification of musical lyrics. The objective of these studies have primarily been to estimate as precise classification models as possible and therefore our project has a quite different focus. Among these studies are the research by Mayer et al. (Rudolf Mayer & Rauber 2008), McKay et al. (Cory McKay & Fujinaga 2010) and several final exam projects from Stanford University, e.g. Sadovsky et al. (Sadovsky & Chen 2006) and Dammann et al. (Dammann & Haugh 2017). In these studies the classification models are extended to include not only lyrics, but also other features, e.g. the model is extended to include not only lyrics, but also audio and album artwork in Dammann et al. In Mayer et al. and McKay et al. they extend their models with ranges of textual features, such as length of text, length of words, characters per word and much more. The decision to include textual features in addition to the standard bag-of-words, in our own classifier, was largely inspired by their experiences.

4 Data

In this section, we give an overview of our data and the data collection procedure employed. Furthermore, we provide a description of the data and consider possible data issues.

4.1 Data source

As our data source we use the website of Genius (genius.com), which is an American digital media company founded in 2009. The content on Genius is edited by an online community consisting of over 2 million contributors and editors, including the artists themselves. Users can share knowledge about artists and musics, furthermore they are able to add, edit and annotate lyrics in the Genius database (Genius Online, accessed 25-08-2018*a*). The strong contributor guidelines and high user activity ensures data quality control, accurate song lyrics and fast lyric updates for new song releases (Genius Online, accessed 25-08-2018*b*).

These qualities make Genius a well fitted source for scraping song lyrics. In our project we collect lyrics within four different music genres being rock, rap, pop and country, because they are popular and give a good representation of different music tastes. Due to the clear labeling of genres the Genius website provides applicable data for both the exploratory analysis and the supervised machine learning on song lyrics. Additionally, we have access to meta data such as release dates, languages and artist names easing the data selection and cleaning process. Finally, Genius has a pagination system that enables a simple collection of song urls.

4.2 Web scraping and quality check

As our data collection tool we use web scraping even though Genius provides a free API. The Genius API only supplies meta data of the songs and not the lyrics, therefore web scraping is necessary for our data collection.

We begin our data collection by identifying how the Genius pagination system is constructed to generate page urls for each genre. Due to an access restriction on pages after page 50 and time considerations, we choose to scrape the first 50 pages in each genre. The scraper iterates through all 200 page urls, to collect the song urls. Each page contains 20 song urls. In total we obtain 4,000 song urls using the python library BeautifulSoup, which amounts to 1,000 song urls from each genre page. However, each song can have multiple genre tags, but only one primary tag. Therefore, there might be overlaps between the songs listed on each genre web page. To avoid the problem of duplicates we collect the song urls in a set. The final set of song urls contains 3,744 unique urls.

Afterwards, the scraper iterates through all song urls to collect data from different parts of the HTML code, using both BeautifulSoup and the re module. The final data extraction for each song consists of the type, language, artist, tag, id, title, lyrics and release date. To ensure a thorough web scraping process we use a function that checks the url multiple times if the get request is unsuccessful. Furthermore, we construct the script to handle unsuccessful request to make sure that the program does not crash. We use a rate limit of one second between each request to not overload the Genius server.

As our main quality check of the web scraping process we use a logging function, which for each url logs whether the data collection was successful. By assessing the log we find that no obstacles occurred during the scraping process. To overcome potential issues regarding non-English lyrics, we selected the English songs using the classification from Genius. Additionally, we remove choir paragraphs from the lyrics, and songs not within our four defined genres. By this exclusion process the final data set is reduced from 3,744 to 3,441 songs. Finally, we format the column headings, reset the index, and create a new column for the release year. The columns of the final data set is 'type', 'language', 'artist', 'tag', 'id', 'title', 'lyrics' and 'release_date'.

Before conducting the analysis we preprocess the lyrics. Firstly, we remove all punctuation exept apostrophes and convert the lyrics to lowercase. Secondly, we tokenize the lyrics using TweetTokenizer, since this library is better at handling words with apostrophes and abrivations.

4.3 Data characteristics

In this section we lay the groundwork for our exploratory analysis by describing the data obtained from the Genius website. One artist can be represented by multiple songs, in our

data set we have 1,024 unique artists, distributed as 259, 294, 243 and 323 across country, pop, rap and rock respectively. Note that the number of unique artists within each genre does not add up to 1,024, indicating that an artists can be unique within multiply genres. The high number of unique artists within each genre indicates that we have a fairly good representation of each genre.

Figure 1(a) illustrates the distribution of song lyrics across the four genres.

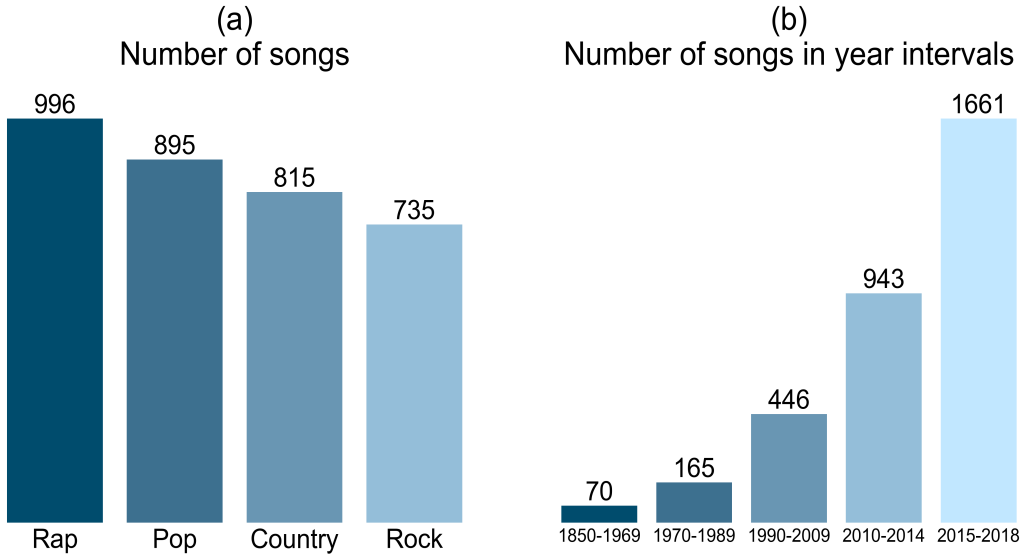


FIGURE 1: SONG DISTRIBUTION *Panel (a) illustrates the distribution of songs across the four genres. Panel (b) illustrates the distribution of songs over five different year intervals.*

From figure 1(a) it is clear that we obtain a good representation across all four genres. Rap is the most represented genre and consists of 994 unique songs, while pop contains the lowest amount of songs with 732. We intended to have an even distribution of songs across each genre, however the tag system of the Genius website have result in overlap between genres resulting in a uneven distribution of songs. Although we have this uneven distribution of songs, each genre is well represented which ensures an applicable data set with clearly labeled data. We will therefore expect that our data can support the identification of differences across the genres.

In figure 1(b), we clearly observe that our data set is dominated by songs with a recent release date, from 2015-2018 1,661 songs were released, while only 70 songs were released in a time period of more than 100 years, 1850-1969. This is due to the ranking system on the Genius website which ranks from the most viewed to the least. As previously explained we scraped the first 50 pages in each genre, which results in a distribution with a bias towards songs with a newer release date. This uneven release distribution also impacts the number of songs in each genre as illustrated in figure 1(a), e.g. one would assume that rap

is more popular today than a century ago. In line with our motivation to analyse what people listen to today, the release distribution of songs fits the purpose of the analysis well. We choose not to remove the older songs, as they have a high view count on the Genius website, thereby enabling them to have some impact on the music listened to today.

4.4 Potential data issues

There are some possible data issues with our data set. Firstly, the Genius website is a user driven community in which users define the primary genre and the second order genres. The definition of the primary genre can cause trouble, because songs nowadays are often produced in collaboration with artists from different genres, e.g. pop songs in which the verse will be more closely related to rap than pop. This can potentially produce difficulties in distinguishing between the different genres and thereby identifying the unique trademarks of the genres, since the genres are not perfectly separated from each other. One way to overcome this would be to enlarge our data set and only include songs with one artist and no featuring artists, however we were not able to use this approach due to the web scraping restriction on the Genius website.

Despite the fact that we use the language detecting tool provided by Genius some of the song lyrics still contain non-English parts. The existence of non-English parts can cause difficulties when comparing the textual features of the songs across the genres. We tried to overcome this problem by leaving out non-English words from the different songs using the langdetect package, however this turned out to be a very unstable procedure. In many cases actual English words and sentences where labelled as non-English. To avoid losing some of the English text in the process, we therefore decide to leave the non-English parts. Since only a handful of the remaining songs contain non-English parts, we do not expect it to be a big problem.

Additionally, lyrics include a lot of slang and abbreviations and thereby the same word can appear in many different forms. This can cause problems with stemming. The stemming tool reduce words to their root, e.g. 'loves' becomes 'love'. However, when words are slang or abbreviated, stemming has difficulty reducing the words, and we risk loosing information. Throughout the analysis, we will therefore test the robustness of our results by using both stemmed and unstemmed lyrics.

Finally, a potential bias in our data set is that the selected songs during our web scraping process are based on the views on the Genius website and went from the most viewed to the least viewed. The views are determined by the users of Genius and might not be an adequate representation of the general population. The users of Genius might have special preferences for a specific type of rap, pop, rock or country, which are not necessarily supported by the broader population.

5 Exploratory analysis

In this section we start our quantitative text analysis by first analysing the language within each genre and performing a comparison between the genres. Afterwards, we investigate the topics of the lyrics. To better understand the differences we perform a topic analysis on each genre. Additionally, we perform a topic analysis on our entire corpus of lyrics to see the topics presented here and if it is possible to classify the genres in different topics. Lastly, we perform a sentiment analysis to see if the sentiment between the genres differ.

5.1 Language

Firstly, figure 2 illustrates the average number of words and the share of unique words in each genre. From figure 2(a) we clearly see that the rap genre is by far the genre that on average contains the largest amount of words, which is partly due to a few exceptionally long rap songs. Country and rock have the lowest amount of words, which is not surprising since instruments usually play a large role and often play solo paragraphs without the company of lyrics. In comparison this is uncommon in rap.

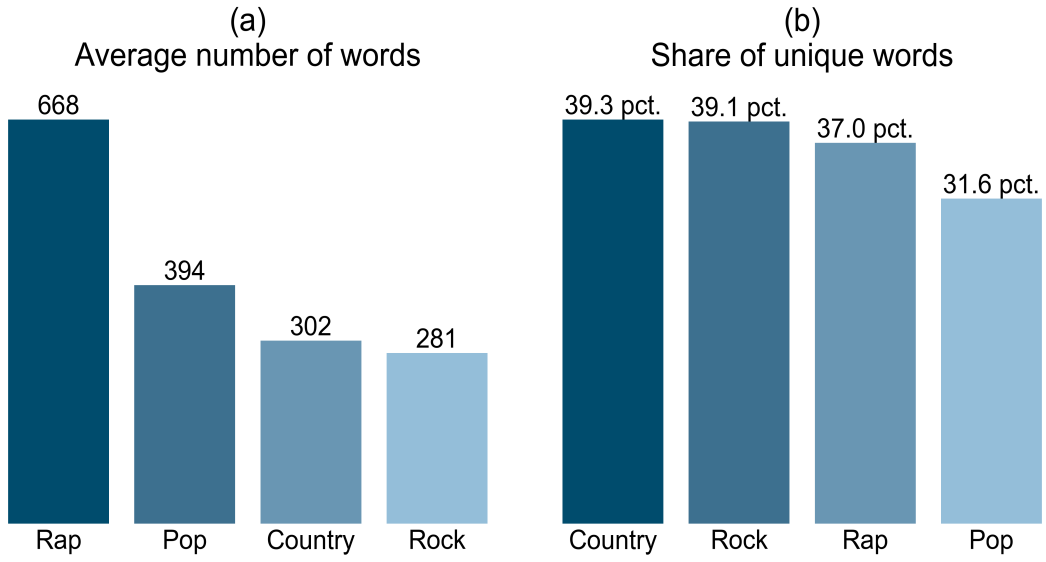


FIGURE 2: WORD DISTRIBUTION *Panel (a) illustrates the average number of words within each genre. Panel (b) illustrates the share of unique words within each genre.*

Figure 2(b) shows that even though the genres country and rock have the lowest amount of words the actual shares of unique words in these genres are largest. This relationship indicates that these genres are not as repetitive as particularly pop, which only contains 31.6 pct. unique words. Even though the pop genre contains the second highest amount of words in each songs, it places last when measured on the amount of unique words. Therefore, this indicates that the vocabulary of pop is not as diverse.

Turning to figure 3 which shows the average characters per word in each genre, we see

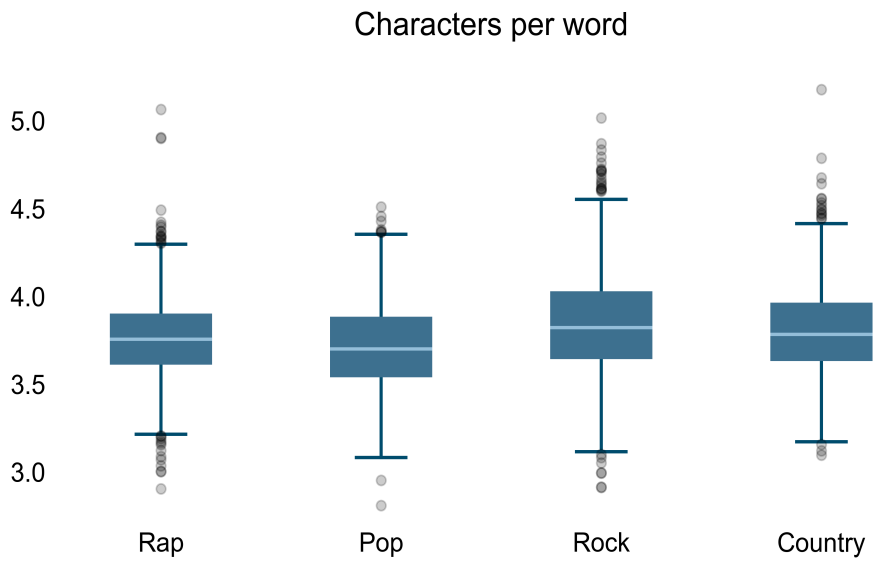


FIGURE 3: BOXPLOT *The figure illustrates the box plot for average number of characters in each word across the four genres.*

that the median of the average word length is balanced around 3.75 for all four genres. Additionally, the four genres cover almost the same range of values. We see that, rock is the genre with most variation in average number of characters per word, whereas rap has the least amount of variation. Pop lyrics are characterized by a slightly shorter average word length than the rest.

In conclusion we find that there exists several differences in language features across the four genres and specifically in areas like word length and share of unique words. This shows that these features might be relevant to include in our prediction model later on.

5.2 Topics

In this section we will explore the different topics that occur within our four genres. Firstly, we intent to show the words that most frequently occur within the different genres. This will be done using word clouds. Additionally, in the following section we perform a topic modelling analysis, which helps us to identify the topics that are present in the different genres.

5.2.1 Word clouds

In order to illustrate the most frequent words in the different genres, we use word clouds. The word clouds are generated based on all songs in each genre. Word clouds are pictures that show the most frequent words in a text, and then scales the words in size based on how frequently they appear. This will help to visualize how the word use differs across the different genres.

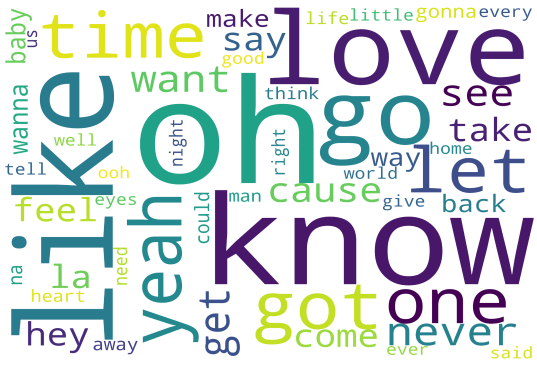
In order to create the word clouds, we create four datasets, one for each genre. After this we create a bag-of-word on each genre. A bag-of-words is a way of extracting features of interest from a text. It consist of two things, the word that appears and a measure of how frequently the word occurs. With a bag-of-words we only consider the frequency of the words, since the order of words in a sentence is not utilized. However, we may face some challenges since in some cases the word order is crucial in regards to the meaning of a specific word. We could have utilized the word order in our bag-of-words by extracting multiple words. This can be done by the use of bigrams or trigrams. However, at this point we are primarily interested in the frequency of the single word to obtain a quick overview of the words use, and will wait to consider bigrams and trigrams in our topic model analysis.

We utilize the CountVectorizer module from sklearn (sklearn Online, accessed 31-08-2018*a*) to construct our bag-of-words. CountVectorizer is a module that counts the number of times each word occurs. In order to overcome potential issues where words without any particular meaning appear. Additionally, we could have utilized a different word counter, particularly the TfidfVectorizer (sklearn Online, accessed 31-08-2018*b*). Further we utilize a stop words parameter, which disregards these kinds of words. Instead of just counting the words this puts little weight on words that appear often and across many lyrics. However here we want a general picture of which words are used within the different genres, why we utilize the CountVectorizer.

The word clouds are shown in figure 4. At a first glimpse of the four word clouds we see that some words appear with a high frequency across all genres. This is words as 'know' and 'like' which are fairly large in all genres. These words are very common and do not necessarily tell us much about content of the lyrics. Additionally, 'sound words' that are very characteristic for lyrics, e.g. 'oh' and 'yeah' also have a high frequency. The appearance of these words indicate that our use of the stop word parameter may not capture all relevant stop words that our data set contains. This is probably because our list is not meant to be used on song lyrics. The stop word package used for our construction of word clouds is created by nltk. However, by analyzing the word clouds we can identify sound words that have not been removed and thereby create our own new list of stop words, which we will return to later in this paper.

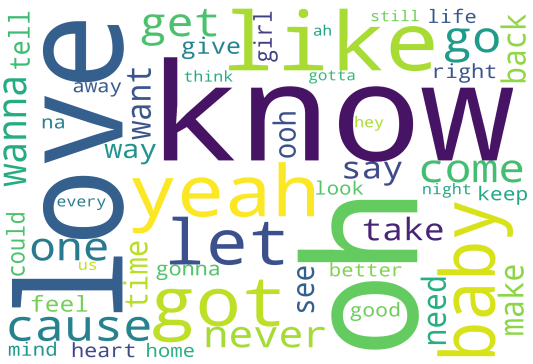
From figure 4, which illustrates the most used words in rock, it is hard to determine any specific words that are unique to rock. However, figure 4 in which the word cloud for the rap genre is shown contain some words that stand out when compared to other genres. Racial words as 'nigga' or swear words like 'fuck' or 'bitch' appear very frequently. Words like these do not appear in top 50 of the most frequently used words in any other genre. This fairly obvious difference between rap and all other genres in regards to the amount of negative word indicates that rap is a genre with a completely different vocabulary and

(A) Wordcloud for rock



(B) Wordcloud for rap

(C) Wordcloud for pop



(D) Wordcloud for country

FIGURE 4: WORDCLOUDS *The figure displays the most used words in each genre by utilizing a word cloud illustration.*

that the theme of the song may not be particularly positive. The two word clouds of pop and country are very similar and share a lot of the same words.

In the word clouds, we observe a very similar tendency in rock, pop and country. Words like 'love' and 'baby' are used often in pop and country and indicates a more positive environment than rap does. It seems very hard to identify unique words when comparing these three genres. Thereby, the only genre that really stands out is rap.

Even though our four word clouds do not identify any significant differences in the usage of words between rock, pop, and country we can use them as an indicator for potential problems we should account for later in this paper when performing our topic analysis. Firstly, a possible way to get better results would be to perform a stemming procedure. Secondly, increasing our list of stop words to also include more specific words that are only used in songs could also help us identify the genres from each other. To further investigate the differences between the genres we perform a topic modelling procedure in the following section.

5.2.2 Topic modelling

Topic modelling is an unsupervised learning method that can assign topics to unlabelled text data. Basically, topic modelling structures unlabelled text by finding common word occurrences and hereby cluster the text into topics (Raschka & Mirjalili 2017).

For this analysis we will use the topic modelling technique Latent Dirichlet Allocation (LDA) with the gensim package (Plus Online, accessed 30-08-2018). Before modelling the topics, we preprocess the lyrics. Firstly, we remove stop words using the nltk package. We know from the word clouds that the lyrics contain many "sound words" such as "oh", "na" and "yeah". As these words will tell us nothing about the topic, but only cloud the results, we extent the list of stop words to include such words as well. Secondly, to improve the performance of the model we use bigrams and trigrams. This will allow the model to overcome some possible problems with polesemy and detect when words belong together as e.g. "new" and "york" in "new_york". Lastly the words are stemmed, so as to transform all words to their root e.g. from "loved" to "love". Stemming can in some cases deteriorate meaning and therefore weaken model predictions. Since lyrics contain slang and deliberate misspellings it is likely that we will loose information when stemming. We therefore run the model on both the original and the stemmed data to compare the results.

In order to investigate the within-genre topics that appear in the lyrics, we begin by splitting the data by genre and running a topic model on each subset. To choose the optimal number of topics, we try different numbers and examine the coherence score, which provides a measure of how good a model is. The results are shown in Table 1.

Table 1: Output for topic modelling with two topics per genre

Genre	Topic	Words
Rap	1	man, right, can't, way, let, wanna, even, em, ass, baby
	2	want, said, money, wanna, still, need, baby, life, man, let
Rock	1	like, know, got, one, want, love, say, i'll, can't, never
	2	love, know, time, go, get, never, like, wanna, come, let
Pop	1	know, love, like, go, baby, cause, wanna, one, let, say
	2	love, like, got, know, come, baby, take, get, cause, never
Country	1	get, got, love, ain't, baby, gonna, cause, little, way, can't
	2	love, back, girl, time, one, take, never, got, go, say

We see that the topics are very unclear, across all genres. Love seems to be a dominant topic in particularly pop, country and rock. Otherwise it is quite difficult to place a label on the topics. The coherence score is generally low for all genres. We find that the coherence score varies only little when changing the number of topics, the boundaries for the filters or when switching from the original to the stemmed words. Generally the

coherence score is a little higher, when we only include few topics. However, as seen in Table 1, even when using only two topics per genre, many of the words are repeated across the topics. This indicates that the number of topics is still to large. Furthermore, the keywords selected are very sensitive to small changes in the filter, e.g. if the 'no_above' option is changed from 0.5 to 0.6 (including words that appear just a little more often in the model) the topics for rap are flooded with new words, particularly "nigga", "bitch" and "fuck" - changing the impression of the topics quite much. All in all, it is clear that the topic model is struggling to segregate the words into meaningful topics within the genres. It could be because there is very little topic variation of the lyrics within each genre, but that is probably not the case. It is more likely that the model struggles to identify the topics because of the size and format of the data. It might be that 700-1000 songs per genre is not enough to capture the more subtle nuances in the lyrics. Also, song lyrics are relatively short texts, with wide usage of slang, spoken language, abbreviations and subtexts, which might make it difficult for the topic model to detect keywords in the text that can be used to "define" the topics.

The topic model has difficulty identifying topics within each genre, but might still be able to identify differences in topics across genres. A different approach to topic modelling is to look at the whole sample of lyrics, set a number of topics, and then calculate the probabilities that songs of a given genre falls under each of the given topics. To conduct this analysis, we split the sample 50-50 into a train and test data set, train the model on the train data and then use the model to place the lyrics from the test data within the different topics. To illustrate, Figure 5 shows the distribution of lyrics across four given topics.

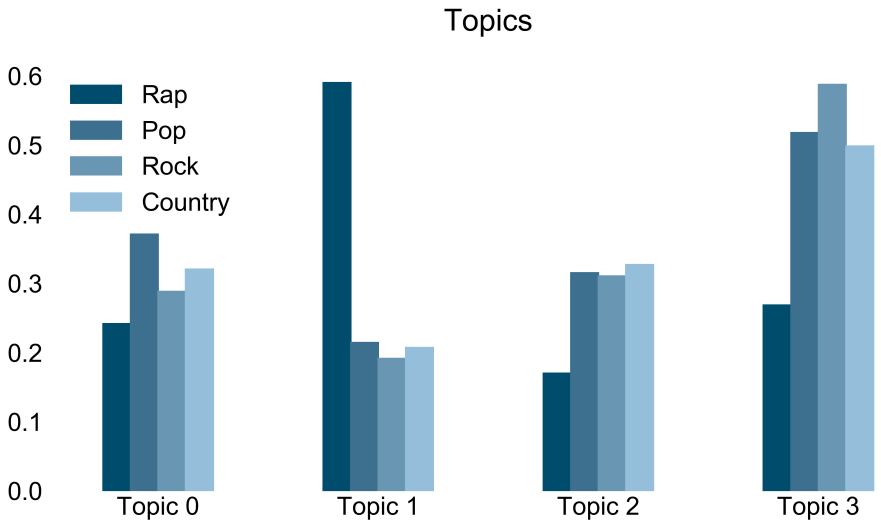


FIGURE 5: FOUR TOPICS *The figure illustrates the distribution of song lyrics across four topics.*

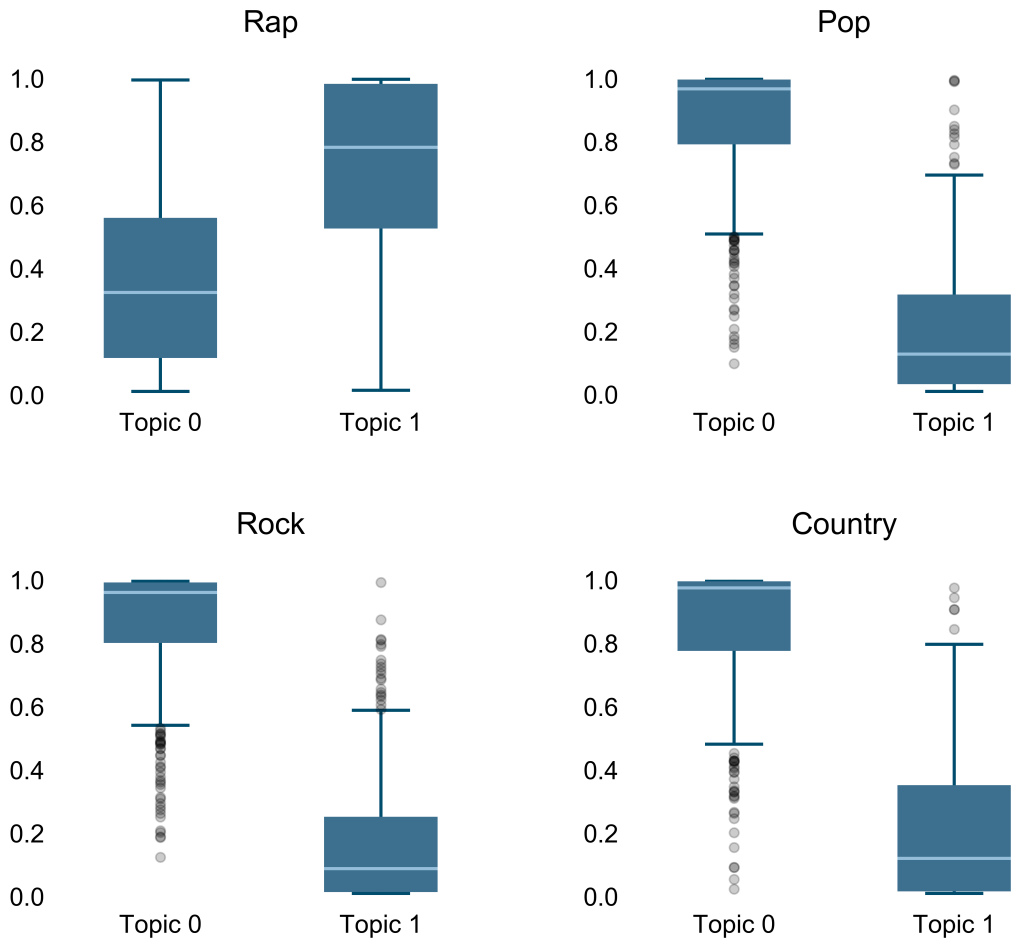


FIGURE 6: BOXPLOTS *The figure displays the distribution of song lyrics across two topics*

If genre was defined solely by the topic of the lyrics. We would expect all rock lyrics to fall within one topic, all pop lyrics to fall within another and so on. In Figure 5 we clearly see that this is not the case. The majority of rock, pop and country lyrics are placed within the same topic, topic 3, whereas the majority of rap lyrics are placed within topic 1. This indicates that the topics of rock, pop and country are very similar, but that the topic of rap is different from the other genres. To further investigate this, we run the model again, but this time on only two topics. The results are shown in Figure 6 and we see the same result as in the case with four topics. Rock, pop and country are primarily placed in topic 0 and rap is primarily placed in topic 1. Our findings suggests that rap distinguish itself from the other three genres, when it comes to topics. As related to the genre-specific topic modelling above, rap was the only genre where we didn't detect a clear focus on love. This might support the conclusion that topics in rap are different.

5.3 Sentiment analysis

In this section we analyse the sentiment of every song and thereby perform a sentiment analysis across the four genres. Our data set does not contain any labels about the sentiment of the songs. Therefore, we use the library VADER to classify the sentiment of the lyrics (cjhutto Online, accessed 28-08-2018). VADER is lexical and rule-based, it matches a sentiment score to each word varying between -4 and 4. Furthermore, the rule-based approach also enables VADER to understand simple negations. We perform a sentiment analysis on each song using the cleaned version of each lyric and by ensuring that all words are lower cased. The output from the VADER package consists of four different scores. Three of the output scores denotes the proportion of the text being either *negative*, *positive* or *neutral*. The fourth score *compound*, is a normalized, weighted composite score of sentiment between -1 and 1. Positive and negative compound score has values above 0.05 or below -0.05 respectively, while neutral songs are found in the interval between -0.05 and 0.05. Figure 7 displays the average positive and negative sentiment composition of all four genres. With 14.7 pct. of the text being negative, the rap genre stands out as the genre with the largest proportion of negative text. In contrast to this, the proportion of negative text in the other genres is centered around 10 pct. with rock being slightly more negative than pop and country. While pop and country are the least negative genres they also account for the largest proportion of positive text, counting 15.1 pct. and 16.3 pct. respectively.

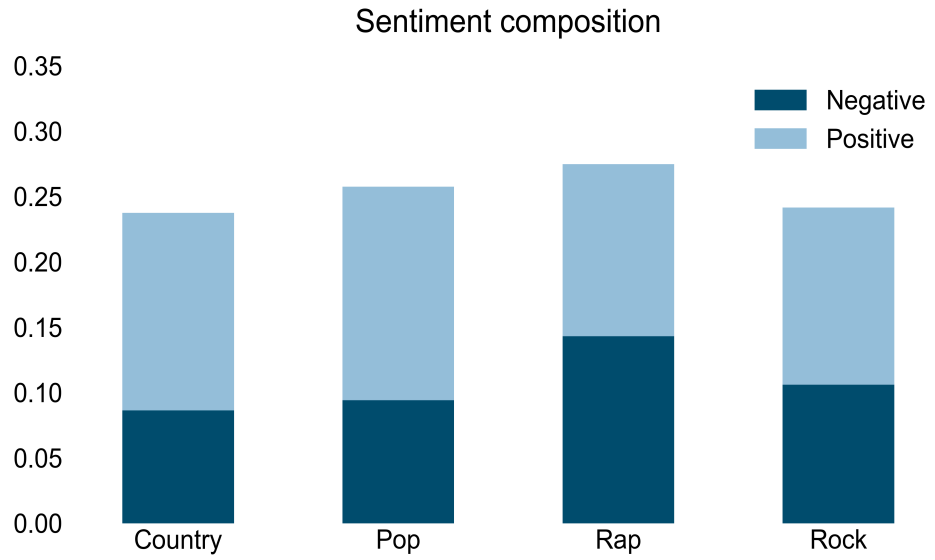


FIGURE 7: SENTIMENT *The figure illustrates the average proportion of text being either positive or negative across all four genres. The proportion of text not displayed in the figure is denoted neutral.*

Turning to the composite score of each category, we have calculated the mean compos-

ite score for each genre. Pop and country has a mean composite score of 0.38 and 0.39, respectively. Hence, this strengthen our picture of pop and country as being genres characterised by positive sentiment. In contrast to pop and country, rap has a mean composite score of -0.18, this supports our view of rap as a genre characterised by a more negative sentiment. Finally, we find that rock is somewhat in between the other genres with a composite score of 0.13. The results of the sentiment analysis indicate as expected that rap usually contains words, which are reflected as being negative while pop and country consists of words associated with a positive mindset.

6 Prediction of genres

In this section, we conduct a proof of concept. To asses the validity of our analysis above we estimate a model to predict the genre of the lyrics. We do this in order to see if the model in fact is capable of distinguishing between the genres, based on the lyrics alone. We take inspiration from Mayer et al. (Rudolf Mayer & Rauber 2008) and estimate a model based on both a bag-of-words and the additional song features investigated in the previous sections. In order to perform the prediction we utilize a Logistic Regression. Despite its name the Logistic Regression is used for classification issues. And is a fairly used method as it is relatively easy to incorporate.

6.1 Creating the Logistic Regression Classifier

There exist many different classification approaches one can take. We choose to apply the logistic regression as this is fairly easy to implement. Additionally the logistic regression takes into account the frequency with which particular words occur (as a bag-of-words). The logistic regression classifier is therefore a good classifier for classifying text documents with different textual characteristics. Therefore, it makes sense to utilize the logistic regression as this approach acknowledges the importance of the different features. The logistic regression classifier calculates a conditional probability of a song belonging to a given genre given its feature vector. The logistic regression is primarily used for binary cases, however our target variable has four outcomes. The logistic regression then utilise a method called one versus rest, where it calculates the conditional probability of a song belonging to genre x in contrast to the rest of the genres(sklearn Online, accessed 31-08-2018*c*). This is done for all genres.

First, we select the relevant features, which we want to include. Our analysis in section 5 indicated that some textual features for the genres differed. Especially the rap genre contained words that were different and we therefore choose to include our bag-of-words as a feature in our classifier. Additionally, we noticed that the number of words as well as the number of unique words were different between the genres. Because of this it seems natural to include these measures as features. Lastly the sentiment analysis showed that

the mood of the song can also have some predictive power, again especially when trying to distinguish rap from the other genres. We only include the positive and the negative sentiment score since the neutral and compound scores can be found from those.

To enable testing of our model we will have to divide our data into a test data set as well as a training data set. We use the training data to fit our model, while the test data is used to evaluate the accuracy of the model on data that it has never seen before. We choose to set our test size to 30 pct. of the overall data set making 70 pct. available for training. Having defined our test and training data we now create our bag-of-words for the songs, we utilize the DataFrameMapper module from sklearn pandas which secures that our bag-of-words can be applied together with the other features. To make our bag-of-words we utilise the CountVectorizer, and remove the most frequent stop words.

Finally, we fit the classifier to our training data as well as transforming it. This ensures that the features in our training data have been standardized. We then apply this transformation to the test features such that they are now comparable.

6.2 Predicting the model

After having standardized and fitted the data sets we are able to estimate our prediction model. We choose a random state number equal to 1, when splitting the data in to test and train data sets, to ensure that our results are reproducible. Initially we fit our logistic regression model, not specifying any hyper-parameters, in order to evaluate how much the model improves from grid-search for the optimal hyper-parameters. Table 2 show the results from estimation of the baseline logistic regression model. The precision of the model is 58.8 pct. which is not particularly high. However, we see from the confusion matrix in table 2 that the model has difficulties with classification of pop, rock and country. The logistic regression model does a decent job in classifying rap from the other genres. The model precision of pop, rock and country varies between 45 and 52 pct., while the model precision of classifying rap is 82 pct. This result is inline with the above analysis, which showed that rap has significantly other characteristics than the other genres. In essence the difference between rap and the other genres, is a probable explanation to why the model is better at classifying rap than the remaining genres.

We now perform a grid-search in order to choose the optimal hyper-parameters being 'C' and 'penalty'. The grid search is performed using a 5-fold cross validation. The parameter 'C' specifies the inverse regularization strength, while 'penalty' specifies the norm used in penalization. Having found and specified the optimal hyper-parameters we estimate the new model. From table 3 we find that the precision of the genre classification increases

	Pop	Rock	Rap	Country	Precision
Pop	89	61	28	45	0.45
Rock	58	134	16	56	0.49
Rap	23	14	252	21	0.82
Country	29	64	10	133	0.52
Total					0.588

TABLE 2: LOGISTIC REGRESSION 1: *The table display the confusion matrix and prediction score for the logistic regression classifier. The feature matrix includes bag-of-words. Hyper-parameters are not optimized, penalty = 'l2' and C = 1.0.*

by 0.5 pct.-points, to 59.3 pct. The precision of classifying the genre rap increases to 85 pct. We see that the model again performs well with classifying the genre of rap, but has difficulties in classifying the other genres from each other.

	Pop	Rock	Rap	Country	Precision
Pop	82	71	24	46	0.54
Rock	58	142	14	50	0.43
Rap	21	14	256	19	0.85
Country	28	68	7	133	0.48
Total					0.593

TABLE 3: LOGISTIC REGRESSION 2: *The table display the confusion matrix and prediction score for the logistic regression classifier. The feature matrix includes bag-of-words. Hyper-parameters are optimized using grid search, penalty = 'l1' and C = 1.0.*

Finally we include additional features from the above analysis. The word count, unique word count, negative sentiment share and positive sentiment share are now included, results from this estimation is reported in table 4. The additional features do not contribute with substantially more predictive power (0.9 pct.-points in precision). In the final model we have an total improvement of precision in classifying pop with 11 pct.-points, while the precision decreases with 6 pct.-points and 4 pct.-points for rock and country respectively. This may be due to that pop differs from the other genres on number of unique words and average number of words, while rock and country is more alike on those two features.

The above results show that the bag-of-words and additional features have predictive power in classification of the genres. Especially rap differs substantially from the other genres, thereby making it easier to classify this genre from the other. With the other genres it becomes more difficult when trying to classify them. In the following section we will go through our discussion, trying to clear what we could have done differently to

	Pop	Rock	Rap	Country	Precision
Pop	92	69	22	40	0.56
Rock	61	142	15	46	0.44
Rap	24	13	257	16	0.86
Country	31	70	5	130	0.48
Total					0.601

TABLE 4: LOGISTIC REGRESSION 3: *The table the display confusion matrix and prediction score for the logistic regression classifier. The feature matrix includes bag-of-words, word count, unique word count, negative sentiment share and positive sentiment share. Hyper-parameters are optimized using grid search, penalty = 'l1' and C = 1.0.*

possible achieve better results.

7 Discussion

While we get reasonable results for the majority of our project, there is still room for improvement in our results, e.g. with regards to the difficulties in distinguishing between the country and pop genres as well as the relatively low prediction accuracy of our machine learning model.

A possible way to ensure a higher rate of distinguishing between the different genres would be to enlarge our data set. Without the restriction on the Genius website we would be able to download more than the tenfold amount that we were actually able to. By using a larger data set we would have a broader range of songs within each genre and thereby a possible pattern may be spotted easier. A larger data set would also enable us to perform a unsupervised topic model and sentiment analysis without the use of predefined packages. In addition to extending our data set, we may also obtain better results regarding the textual analysis by improving the reduction of stop words even more.

In creating our bag-of-words we utilise the CountVectorizer from sklearn. As we have already covered there might exist other approaches to creating bag-of-words as for example the TfidfVectorizer from sklearn. It could be interesting to see how the logistic regression model performed using this bag-of-words approach instead, and this could be a point for further analysis.

Additionally, there is no guarantee that the logistic regression method is the best classifier to handle our particular problem. Other approaches such as Decision trees or Random forest classifiers might yield better results. For further analysis one might consider to utilise these approaches to see if it is possible to obtain better prediction results.

Of course, there also exists other textual features that one might consider to extend

the classification model with. As mentioned in the literature review Cory McKay & Fujinaga (2010) describes different textual features that one might consider including in a prediction model for song lyrics. We have included some of their suggestions, but it could be interesting to also include even more, e.g. some variance features, that they have shown to have predictive power. The variance feature could have been particularly interesting to include in our predictive model as we saw in our analysis that the variance for some of the textual features also varied across different genres.

8 Concluding remarks

In conclusion, we find compelling evidence of different textual features being present across our four investigated genres. Features as word length and share of unique words appear to be important determinants of what kind of genre the song belongs to. However, we do not find any support of a clear distinguishing between pop, country and rock and only rap turns out to be different, when utilizing word clouds and topic modelling. The sentiment analysis performed supports the findings of our topic analysis, since the rap genre seems to consist of a more negative words than the other genres. Furthermore, by using a logistic regression to predict the genres we obtain a very high precision of rap and mediocre for pop, country and rock. Lastly, by including the results of our exploratory analysis we achieve an even higher overall precision, thus we argue for including the findings of this analysis in future genre classification.

9 Bibliography

References

cjhutto (Online, accessed 28-08-2018), 'Vader-sentiment-analysis', `https://github.com/cjhutto/vaderSentiment`.

Cory McKay, John Ashley Burgoyne, J. H. J. B. L. S. G. V. & Fujinaga, I. (2010), 'Evaluation the genre classification performance of lyral feature relative to audio, symbolic and cultural features', *CIRMMT* .

Dammann, T. & Haugh, K. (2017), 'Genre classification of spotify songs using lyrics, audio previews, and album artwork', *CS229 Final Project* .

Forbes (Online, accessed 31-08-2018), 'Americans are spending more time listening to music than ever before', `https://www.forbes.com/sites/hughmcintyre/2017/11/09/americans-are-spending-more-time-listening-to-music-than-ever-before/#7ddb784f2f7f`.

Genius (Online, accessed 25-08-2018*a*), 'About genius', `https://genius.com/Genius-about-genius-annotated`.

Genius (Online, accessed 25-08-2018*b*), 'How to add songs to genius', `https://genius.com/Genius-how-to-add-songs-to-genius-annotated`.

Marshall, S. R. & Naumann, L. P. (2018), 'What's your favorite music? music preferences cue racial identity', *Journal of Reasearch in Personality* **76**, 74–91.

Plus, M. L. (Online, accessed 30-08-2018), 'Topic modelling with gensim', `https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/`.

Raschka, S. & Mirjalili, V. (2017), *Python machine learning*, Packt Publishing Ltd.

Rudolf Mayer, R. N. & Rauber, A. (2008), 'Rhyme and style features for musical genre classification by song lyrics', *ISMIR* .

Sadovsky, A. & Chen, X. (2006), 'Song genre and artist classification via supervised learning from lyrics', *CS224N Final Project* .

sklearn (Online, accessed 31-08-2018*a*), 'sklearn.feature$_e$*xtraction.text.countvectorizer*', .

sklearn (Online, accessed 31-08-2018*b*), 'sklearn.feature$_e$*xtraction.text.tfidfvectorizer*', .

sklearn (Online, accessed 31-08-2018*c*), 'sklearn.linear$_m$*odel.logisticregression*'.