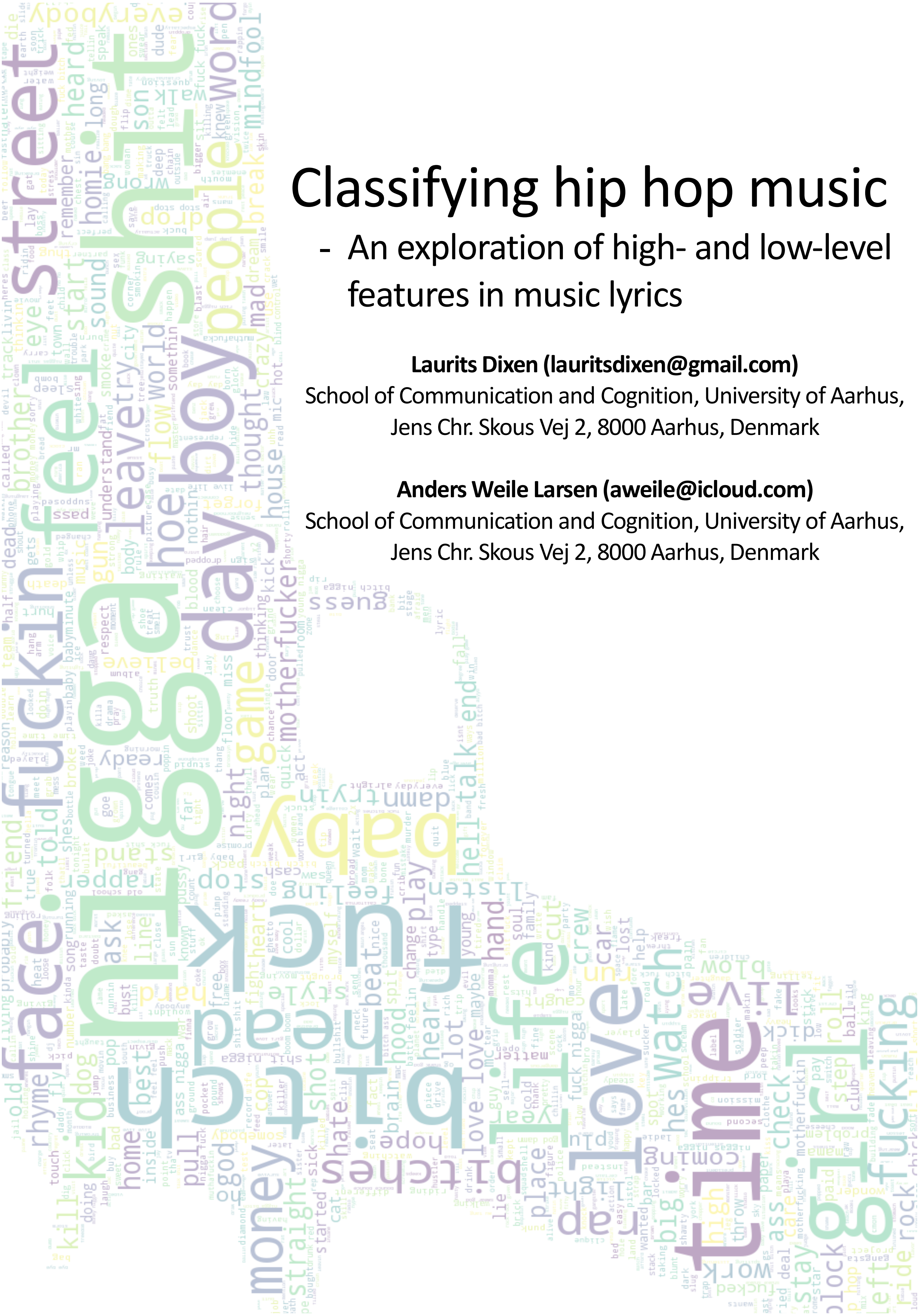

Classifying hip hop music

- An exploration of high- and low-level features in music lyrics

Laurits Dixen (lauritsdixen@gmail.com)
School of Communication and Cognition, University of Aarhus,
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Anders Weile Larsen (aweile@icloud.com)
School of Communication and Cognition, University of Aarhus,
Jens Chr. Skous Vej 2, 8000 Aarhus, Denmark

Abstract

This study aims to investigate the nature of music genres and attempts to classify songs appropriately utilizing statistical tools. Specifically, we tested whether hip hop songs differ enough from other genres to be identified using only music lyrics. With a dataset of 50,000 hip hop song lyrics and 65,000 pop and rock song lyrics, we achieved over 90% accuracy in our highest performing model which suggests hip hop lyrics can be seen as a taxonomic class. A distinction between low- and high-level features is introduced and models applying one or the other is compared. As low-level features, we relied on six different text statistic features such as the number of lines and the mean length of words in a text. For a high-level feature, we employed a probabilistic topic model known as Latent Dirichlet Allocation. The low-level model (87.67% accuracy) performed better than the high-level model (85.85% accuracy) which suggests that non-meaningful features are more indicative to hip hop lyrics than meaningful features. However, a model with both feature types combined yielded the best results (90.29% accuracy). We regard these results as moderately successful and suggest different ways to improve the model.

Keywords: Music information retrieval, automatic classification, music lyrics, topic modelling.

* The main author of each paragraph is specified after the title of the paragraph in brackets with "L" standing for Laurits Dixen and "A" standing for Anders Weile Larsen. We want to stress that each author contributed to all sections of the paper. In cases where both authors were near equally responsible, "L, A" was put in brackets after the paragraph's title.

Table of Contents

Introduction (A)

Defining music genres is difficult. First of all, music is a combination of internal factors such as audio and lyric content and external factors such as culture and setting. Secondly, individuals' experiences of music are different and change over time. Since we have to conceive that the ways in which we categorize music are on some level derived from how we experience it, we intuitively expect genre boundaries to be vague. In spite of this, it would seem absurd to claim that genre boundaries are completely arbitrary, just as it would seem absurd to claim that genre boundaries are impermeable.

Mendoza-Halliday (2017) presents three theoretical frameworks for thinking about music genres. He states that one can think of genres as cognitive categories, taxonomic classes, and as cultural units. In the framework of cognitive categories genres are situated, given that perception of genres rely on the meaning mapping structure of human brains. The idea is that categories are determined by resemblances rather than set properties which is the case in taxonomic genre theories. In the framework of taxonomic classes, genres are sets with essential properties. In order for a song to be a member of a specific genre, the song must contain these features that characterize the genre. In fuzzy set theory membership of a set is not absolute, but a scalable measure. As such, sets, in this case genres, have so-called fuzzy boundaries and cannot be defined unequivocally. The cultural unit framework emphasizes genres as being normalized through social interactions. This means that genre conventions are established implicitly via intersubjective alignment.

Whereas different genre theorists often cling to one or the other of these genre theories, Mendoza-Halliday (2017) proposes that the theories are not irreconcilable, but should be viewed as conceptual states in a cycle. He suggests that social conventions mediate the transition of a genre from a cognitive category to a taxonomic class. When a music genre has been determined as a taxonomic class, the mere thinking about the genre will restart the cycle, since the genre in question once again is being subjected to the thinking person's cognitive dispositions. While cognitive categories and cultural units are context-dependent and cannot be fully explained by fuzzy logic, taxonomic classes do not depend on context and can be fully explained by fuzzy logic.

Following Mendoza-Halliday (2017), one cannot simply think about a particular genre, subgenre, or group of genres as a monolithic entity, no matter how well-established, because it is constantly subjected to revisions. It also implies that we can make the assumption of taxonomic organization,

at least to a certain extent. If we can locate elements that are members of a certain genre, we would be able to systematically predict whether any particular song is a member of this genre based on the particular elements of the song. This is exactly what this study aims to do: to examine whether it is possible to classify genres based on linguistic features in the framework of genres as taxonomic classes.

More specifically we attempt to utilize modern statistical methods to classify hip hop songs solely by extracting features from their lyrical content. Hip hop was chosen because of the genre's very distinguishable lexical and morphological features rooted in African American Vernacular English as well as its structural and thematic properties that are also strongly rooted in African American culture (Alim, 2004). Because hip hop is so strongly influenced by the culture in which it was conceived, it is very likely that one is able to find several distinguishable characteristics of the genre. On the basis of statistical analyses on musicians' profiles on Myspace, hip hop has been found to have strong boundaries to other music genres as well as little differentiation within the genre (Silver, Lee, & Childress, 2016). This makes hip hop ideal for this type of study because it increases the validity of assumptions of the insignificance of context and genre-feature taxonomy. This simply means we assume that hip hop as a lyric genre subdivides into properties which can be examined via statistical methods without further qualitative investigations. We call these properties lyrical features.

A previous study on genre classification by song lyrics implemented a distinction of lyric features into text statistic features, part-of-speech features, rhyme features, and bag-of-words features (Mayer, Neumayer, & Rauber, 2008). One could view this as a hierarchical structure which increases in complexity as we move from simple text-statistic features to more complex bag-of-words features. These subcategories enable us to ask questions about the nature of hip hop as a lyric genre like: "What kind of lyric features have a bearing on hip hop?" In order to connect these kinds of questions to a larger discussion in the field of semiotics, we could ask questions like: "Are meaningful features less or more essential than features that carry no inherent meaning?" Without disregarding the lack of consensus on the exact nature of meaning, we adopt a definition of meaningfulness that corresponds to the aforementioned high- and low-level structure of song lyrics. To make the distinction clear and limit the area of study we choose to examine only text statistic features and bag-of-words features. Bag-of-words features are considered meaningful while text

statistic features are considered less meaningful. This is based on the assumption that statistical modeling based on bag-of-words are more similar to how we humans consciously perceive and talk about a song's lyrical contents than text statistics.

In addition to a model containing both low-level (text statistic) features and high-level (bag-of-words) features we construct two submodels: one containing only low-level features and one containing only high-level features. If our models are proven successful, they should be able to inform us about the significance of both low-level and high-level features. Significance will be tested as each set of features' abilities to correctly categorize a corpus of song lyrics. The additional subdivision of features into low-level and high-level features calls for an ordinal structuring of our research questions:

1. Are we able to computationally classify a song as hip hop or non-hip hop based solely on linguistic features?
2. Are low-level linguistic features more or less essential than high-level linguistic features when classifying songs into hip hop and non-hip hop categories?

Low-level features (A)

Text statistic features are defined as features that are based solely on counting the occurrences of generic features at different levels in each song's lyrics. These levels are characters, words, and lines, and the features do not require any supervision. As many of these generic features are influenced heavily by the total length of a song's lyrics, these are normalized by dividing with the total number of words in each song text. Inspired by Mayer et al. (2008) we apply six such features measured for each song's lyrics listed below:

Feature name	**Operationalization**
Number of words	The total number of words.
Number of unique words	Total number of unique words divided by the total number of words.
Number of lines	Total number of individual lines.
Words per line	Total number of words divided by the total number of lines.

Length of words	The average length of each word.
Number of punctuation marks	The number colons, single-quotes, commas, hyphens, semicolons, question marks, and dots divided by the total number of words.
Number of choruses	The number of choruses divided by the total number of words.

The idea with implementing these features is to capture the idiosyncratic properties of both the overall structure of hip hop lyrics and the internal word structure within hip hop lyrics. A large part of most hip hop songs' lyrics consists of rap, and the terms 'hip hop' and 'rap' are often used synonymously to refer to the genre as a music genre. Rap distinguishes itself from other types of song lyrics by being based largely on rhythm rather than melody (Attridge, 1995). The fast-paced rapping parts of the hip hop songs are often found in the verses of the songs whereas choruses are frequently sung (Anbari, 2010). Therefore, we expect that the number of choruses per total amount of words in hip hop is significantly lower than in the other genres and that the number of words and the number of lines are considerably larger for hip hop than for other genres. Since rapping involves high levels of creative slang and wordplay (Edwards, 2009), we would expect the number of unique words to be relatively high for hip hop songs. The unique spelling in rap (Olivo, 2001) entails an increased number of punctuation marks (e.g. "'stead" and "slow-mo" instead of "instead" and "slow motion"). We do not hypothesize any certain prevalence for the mean length of words and words per line, but speculate that there might be effects. It could be that hip hop lyrics have longer words and/or more words per line and vice versa.

High-level features (L)

High-level features are meant to capture the semantics of a text in some meaningful way. A high-level feature goes beyond the simple structural aspects of the low-level features. Actually, a high-level feature model often actively removes some of these low-level features on purpose to better understand the semantics of the text. The high-level feature model approaches the document as a whole and attempts to assess what the lyric (document) is *about*, or what feelings it expresses. A disadvantage for high-level features is that some human interaction is necessary for the method to

be effective (Blei, 2012). But this will in turn give us features which are easier to interpret for humans.

As our high-level feature model, we will primarily be working with a topic model called Latent Dirichlet Association (LDA) as introduced by Blei, Ng. & Jordan (2002). This model attempts to assign a topic to a document based on its relation to other documents like it. We also experimented with sentiment score analysis (Volcanu, Yanon; & Fogel, David B, 2001), which aims to score a text on various scales of sentiment for example positive/negative or dominating/submissive. In summary: this study utilizes topic models and sentiment scores as high-level features. These methods will be explained in further detail later.

A bag of words (L)

To interpret the text holistically, which a high-level feature model does, is extremely difficult, and some preprocessing and reduction of data is necessary. The key assumption in our model is often referred to as 'bag-of-words' (Wallach, 2006). A bag-of-words model refers to a model that metaphorically takes all the words in a text, puts them in a bag and shakes them around. This means all syntax is thrown away and only word counts remain.

In his article 'Distributional Structure' Zellig S. Harris (1954) discusses the possibility that language has an underlying (statistical) structure or distribution. In other words, whether certain words or phonemes regularly appear together in clusters or not. Harris (1954, p. 156) mentions that "language is not merely a bag of words", but also points out that structures and formal distributions are present in all corners of language. This results in what is now referred to as the distributional hypothesis, which is stated something like this: "words in similar context have similar meaning" (Sahlgren, 2008). This hypothesis makes it possible to operationalize something as vague as meaning, and as we move along we will see how this framework will be very fruitful for computational modeling of language.

The bag-of-words assumption also clearly has some downsides. The semantic content of a word or a sentence is strongly influenced by the context it appears in, which is shaken away in our metaphorical bag. A sentence like 'The dog bites the man' clearly is different from 'The man bites the dog', but this difference is not represented in the bag of words, as the order of the words is

thrown away. The model will also struggle on more abstract concepts that are heavily influenced by context like irony, humour, and metaphors.

Word vectorizations and document-term matrices (L)

As words are not inherently meaningful for computers, we need to implement a way of turning a text into a set of numbers, i.e. operationalizing the distributional structure. For this, word vectorization is a very popular set of methods (T. Mikolov et al., 2013). These methods all have different computational or analytical advantages and are mainly split up into count-based and predictive methods (M Baroni, G Dinu, G Kruszewski, 2013). For this study, we will focus on simple count-based methods because it enables us to perform our chosen high-level feature models (LDA and sentiment score).

The idea behind count-based word vectorization is modest but powerful. We simply count the number of times a certain word appears in a text and then create a vector for it. To illustrate: consider a case in which we have a corpus of text consisting of three documents:

'I love hip-hop' 'I love music' 'I hate pop'

Here, a document is only a sentence, but it could be any length of text that in some way belongs together. The first document will have the values: {I:1, love:1, hip-hop:1}, the second will have {I:1, love:1, music:1} and the third: {I:1, hate:1, pop:1}. Each of these sentences has now been expressed as a three-dimensional vector, that we can plot in a three-dimensional space and perform algebra on. Of course, these vectors need to be represented in the *same* dimensions before we can compare them. This is why we will introduce the document-term matrix (DTM), a matrix of all words in all documents. This is a way of combining multiple word vectors into a single framework. The case-example will result in the following DTM:

	I	Love	Hate	Hip-hop	Music	Pop
Document 1	1	1	0	1	0	0
Document 2	1	1	0	0	1	0

Document 3	1	0	1	0	0	1

We can now look at the distances between these vectors and get an idea of which are closer to each other semantically, or we can ascribe values to each word and sum them up if we want to compare them on a specific axis.

Sentiment analysis (L)

Word vectorization, as illustrated above, enables us to perform sentiment analysis (Volcani et al., 2006). In sentiment analysis each word is scored on a scale (positive/negative), then the average score for each document can be compared to get an idea of which document is more positive or negative. In our example, a word like 'love' is probably scored more positive than 'hate', and document 3 will then have a lower sentiment score than the others. This method has wide-reaching implications (Pang & Lee, 2008), but heavily relies on previously scored word-sentiment corpuses. For the method to be reliable, the scoring must be independent of the documents but in accordance with the context, the documents appear in (Pang & Lee, 2008).

Although Liang, Gu, and O'Connor (2011) reported sentiment analysis to be of no value when classifying several genres at a time, we anticipate that sentiment scores will prove significant in a binary genre division between hip hop and non-hip hop. We expect hip hop lyrics to be generally more negatively laden than non-hip hop lyrics.

Latent Dirichlet Allocation (L)

In 2002, Blei, Ng. & Jordan introduced a new form of probabilistic topic model called Latent Dirichlet Allocation (LDA), which builds on top of Thomas Hofmann's (1999) latent semantic analysis. A probabilistic topic model is an algorithm that tries to reveal hidden layers in a large set of data using advanced Bayesian statistics (Blei, 2012). It is conventional to refer to these hidden layers as topics (hence topic model), but in this study they also will refer to genres, as we attempt to match the hidden layers to different genres. Although topic models can be used on various forms of data (Blei, 2012), we will exclusively be discussing its applications for text data. It is also worth mentioning that

LDA is an unsupervised algorithm. That is, it works on unlabeled data, and it will therefore not utilize the labels in our dataset.

LDA is one of the simpler probabilistic topic models. It works by trying to identify a few keywords for each topic, then it allocates a topic to each document based on the appearances of these words. The keywords are words which appear frequently in one topic but not in others. This means that the keywords are particular to one specific topic. A topic is thus defined as the probabilities for each keyword (Blei, 2012). This is best described by what is called *the generative process*, which is an assumption the model has to make before it has seen the text. The idea is, that each topic is represented in the data and that all documents exhibit some amount of that topic. Then it assumes that when a document was written the author decided on the distribution of topics, and then just chose random words from topics based on this distribution (Blei, 2012). This sounds like a highly unrealistic assumption, but remember our distributional hypothesis of a hidden semantic structure, which LDA is working towards revealing. LDA is clearly never going to produce any meaningful text with this assumption as, like any other model, it simplifies reality greatly, but it will identify meaningful topics for a corpus.

Another key assumption behind LDA in specific is that while every topic is present in all documents, only a few are significant (Blei, 2012). Therefore, LDA will minimize the number of topics per document, but it also tries to maximize the probability of a few words in each topic. The two goals are conflicting, as it is hard to assign topics when only a few words are indicative of each topic. Through an iterative process, LDA will try to balance them out (Blei, Ng. & Jordan 2002). To sum up: LDA is trying to identify a small number of words for each topic that clearly expresses what the topic is about. For example, in a love song the words: 'is' and 'she' are probably very common, but they are not informative as to which topic the song belongs to. This is because the words 'is' and 'she' appears just as frequent in other topics as much. Depending on what other topics that are present in the corpus, LDA will find the words with the most love-ness. That could be 'love', 'face', 'eyes' etc., which presumably would not be likely to see in a song about street life for example. Hopefully, these words can now easily be interpreted by a human and thereby categorize a particular document.

Materials and methods

Our golden standard for genre - Wikipedia.org (L)

Collecting data and deciding how to classify it played a crucial role in this project. Finding a dataset with predefined genres is not easy, as genre definition change from person to person. With this in mind, we decided on using Wikipedia as our golden standard for genre, adhering to a sort of 'wisdom of the crowd'-principle. Wikipedia keeps lists of all their articles and which category they belong to. This categorization was our golden standard so that all artists named in e.g. ("List of hip hop groups", 2017) were defined as hip hop artists. Using this method, 1800 hip hop artist names were collected. For our contrast to hip hop music, we decided on using pop and alternative rock. Both of these genres are very broad and will therefore represent music more generally. We ended up with the names of 2000 pop artist and 1500 alternative rock band on our list. A full list of the Wikipedia lists used in this study, can be found in the appendix.

Collecting lyric data - Genius.com (L)

After having around 5300 artist names and their associated genre, we turned to Genius.com for the actual lyrics. Genius (former RapGenius) is a crowdsourced website for annotating lyrics. Only recently it was changed from exclusively rap lyrics to all genres. Therefore, it has a much broader selection of hip hop lyrics than other genres, which balanced out the fewer number of hip hop artists. Genius has listed popular albums for an artist and the 10 most popular songs that artist has been a part of, and all of these were collected under the artist's genre. The hip hop songs were collected first, which means if a song featured a hip hop artist in any way, it would be classified as hip hop, e.g. if Eminem (a hip hop artists) is featuring on a non-hip hop song, it would still be classified as hip hop. This is of course a debatable decision and could easily be done in a different way. After filtering out duplicates we ended up a total of 115 000 song lyrics: 50 000 from hip hop, 31 000 from pop and 34 000 from alternative rock.

Cleaning data (L)

Lyrics from Genius.com is not formatted as clean lyrics. It includes some amount of metatext of the song, including the artist, what parts of the lyrics are chorus and verse and repetitions (e.g. 2x). These data are mostly irrelevant when working with the text and were removed, along with punctuations, in the clean version of the data. Because of the far-reaching method we used for collecting the lyrics, we also obtained non-English lyrics, which needed to be filtered off. Using a normal English corpus would not work, because many of the texts were only part non-English, and hip hop lyrics use many different slang words. For example, verbs in present participle are very often written without the 'g' in the end (e.g. trying/tryin, running/runnin, often runnin'). After adding common English slang words from hip hop (e.g. 'yall', 'murda', 'dat') and onomatopes common used in songs (e.g. 'woop' and 'ohh'), we gave each song a score of how many percents of the words were in our corpus. All songs with under 80% were sorted out as non-English, which may seem high, but these songs were presumably not well suited to be analyzed anyway. Furthermore, songs with lyrics shorter than 100 characters, and songs that contained the sentence 'lyrics for this song is not yet transcribed' were removed. Finally, some artists on our list had clearly nothing to with music lyrics (e.g. William Shakespeare, Game of Thrones, and C.I.A.) and were removed as well. After all the cleaning we were left with 105 000 songs in total: 45 000 hip hop and 60 000 non-hip hop.

Operationalizing high-level features (L)

Constructing the LDA-model and the sentiment scores required a few preprocessing steps. First, LDA utilizes stop words, words that are so common in language that there is no insight gained from highlighting them. Usually, stop words are chosen as the most common words in a language (e.g. 'the' or 'a'), but in the case of music lyrics, few typical stop words are present. Furthermore, hip hop artists are known for their slang, and it could be that a defining characteristic of a non-hip hop song is precisely the use of common English words. Because of this, we created a custom list of stop words, consisting of words that appeared with the same frequency across genres. You could call them music stop words, words that appear frequently in all types of songs. We also filtered out all words only occurring in 5 or fewer documents. Then the model ran on the 10 000 most frequent

words in the data with only 2 latent topics, presumably hip hop and non-hip hop. The python module 'SciKit Learn' was used for both vectorization and training the model (Pedregosa et al, 2011).

All this was done on 10% of the complete data. A test-train data split of 90-10 was necessary to avoid overfitting the data, but not detrimental to performance due to a large dataset. It is important to point out the vectorization of the test data was done on the trained vector. That is, every unique word in the training set represents a feature and no new features can be introduced during training. For training, we introduced a feature called 'Affiliation'. This is a number between 0 and 1 for every topic, representing the affiliation to that topic. In the generative framework, it is how influential that topic was in creating the document, it can be thought of as how much of the lyrics that can be classified as hip hop. The test is simply if the affiliation level is above 0.5 for the topic that is hip hop. Note that it is necessary that a decides which topic is hip hop and which is not. LDAs with more than 2 topics was tested as well, but none performed better than the binary version. We avoided stemming in the LDA, as slang words were handled very poorly in available implementations.

For the sentiment analysis, the general-purpose 'Harvard-IV dictionary' ascribed each song lyrics in the data a sentiment score using the package 'Dictionary-Based Sentiment Analysis' (Feuerriegel & Proellochs, 2013) in RStudio (RStudio Team, 2015). This package stems (reduce to the word stem) words before scoring them. Individual words contained in the Harvard-IV dictionary are ascribed a valence value, either positive or negative. The sentiment score for each song lyrics is then calculated from the relative occurrences of words that are deemed positive and words that are deemed negative by the dictionary. The sentiment scores were kept as continuous variables going into the classification model and not converted into binary or otherwise categorical variables as we would expect our classifier to perform better if more information is provided.

Constructing models (A)

Low-level features were coded, and the genre of each song text was changed into a binary genre category so that each text would be classified as either hip hop and non-hip hop. Afterwards, the data was split into a training set consisting of 80% of the untrained data and a test set consisting of 20% of the untrained data. In order to randomise the distribution of data into the two sets yet maintain consistency in the training and test data, we made use of a fixed seed.

Our models are evaluated by their ability to correctly categorize an unknown set of song lyrics as either hip hop or non-hip hop. This binary definition of genre is the outcome variable of our models. Since the outcome variable is binary, logistic regression can be used as a simple classification algorithm with the features acting as predictor variables. For each song lyric, a logistic regression model can calculate the probability of it being a hip hop song. This probability is synonymous with the concept of degree of membership in fuzzy set theory. If it is more than 50%, we predict that it is a hip hop song, and if it is less than 50%, we categorize it as non-hip hop. In order to answer our research questions, we developed three models: a model using low-level features as predictor variables, a model using high-level features as the predictor variable, and a combined model using both low-level and high-level features as predictor variables. Conducting multiple logistic regression with non-hip hop as baseline category, the three models were defined[1]:

Combined model: Genre ~ nWords + nUniqueWords + nLines + wordLength + nPunctuation + nChoruses + affiliation + sentiment

Low level model: Genre ~ nWords + nUniqueWords + nLines + wordLength + nPunctuation + nChoruses

High level model: Genre ~ affiliation + sentiment

Before finalizing the models, a correlation test was performed on the combined model in order to detect any covariance between low-level features and high-level features. Since we want to compare the high-level model and low-level model, any significant correlation between high-level features and low-level features would be problematic. The threshold for removing predictor variables was set to ±.3 indicating a moderate correlation. For each of these three models accuracy scores, positive predictive values (PPVs) and negative predictive values (NPVs) were calculated. The train/test split, logistic regression, and model performance measures were conducted in RStudio (RStudio Team, 2015), and low-level features were either coded in Python or in RStudio (RStudio Team, 2015).

[1] 'n' stands for 'number of'. E.g. 'nWords is the number of words.

Results (A)

Combined model

The combined model containing all high-level and low-level features was shown to significantly predict genre: $X^2(7) = 59{,}391.52$, $p < .001$. An assessment of the individual predictor variables revealed them to all be significant (all *p-values* < .001) except for 'words per line' and sentiment scores (both *p*s > .05) which were subsequently removed from the model. The contributions of the individual predictor variables are displayed in the following table:

Stats / Variables	*b*	*SE*	*z*
Number of words	9.43e-03	9.43e-03	60.00***
Number of unique words	4.27	0.12	35.95***
Number of lines	-1.11e-2	9.24e-4	-12.06***
Length of words	-0.33	4.60e-2	-7.09***
Number of punctuation marks	2.10	0.16	13.44***
Number of choruses	-4.46	-2.20	2.75e-2*
Affiliation	4.22	4.77e-2	88.36***

* $p < .05$, *** $p < .001$

The accuracy measures for the model were:

Accuracy (95% CI)	**PPV**	**NPV**
90.29% (89.86% – 90.71%)	91.29%	89.60%

Since no correlation coefficient values between high-level features and low-level features exceeded the threshold of ±.3, none of the predictor variables were removed from neither the low-level nor the high-level model.

Low-level model

The low-level model significantly predicted genre: $X^2(6) = 50{,}259.76$, $p < .001$. An examination of the individual predictor variables revealed them to all be significant (all *p*s < .001) except for 'words per

line' ($p > .05$) which was subsequently removed from the model. The contributions of the individual predictor variables are displayed in the following table:

Stats / Variables	*b*	*SE*	*z*
Number of words	1.43e-2	1.49e-4	95.75***
Number of unique words	7.95	0.11	73.12***
Number of lines	-1.47e-2	8.64e-4	-16.99***
Length of words	-5.27e-1	4.36e-2	-12.08***
Number of punctuation marks	5.81	0.14	40.43***
Number of choruses	-23.50	1.99	-11.79***

*** $p < .001$

The accuracy measures for the model were:

Accuracy (95% CI)	**PPV**	**NPV**
87.67% (87.2% – 88.14%)	89.76%	86.34%

High-level model

The high-level model was also shown to significantly predict genre: $X^2(1) = 49{,}067.72$, $p < .001$. Sentiment scores were found to be insignificant ($p > .05$) and were removed from the model. The hip hop affiliation measure had a significant impact on genre: $b = 6.50$ ($SE = 0.042$), $z = 156.1$, $p < .001$. The accuracy measures for the model were:

Accuracy (95% CI)	**PPV**	**NPV**
85.85% (85.34% – 86.34%)	89.59%	85.36%

Discussion (L, A)

The results show that high levels of accuracy for hip hop classification can be achieved based on linguistic features of lyrics. Including seven features, six low-level and one high-level, nine out of ten songs were correctly classified. Furthermore, the low-level model had an edge in accuracy measures when compared to the high-level model. For all three models, the PPVs were larger than the

respective NPVs showing that the classifiers were more likely correctly classify hip songs than they were at correctly classifying non-hip hop songs. The number of false negatives were proportionally larger than the number of false positives. A qualitative post-hoc analysis of wrongly classified songs revealed one distinct pattern: False negatives with very low predicted degrees of membership to hip hop often had a very low number of total words and unique words, and false positives with very high predicted degrees of membership to hip hop were often very long pop songs. Crucially, it seemed as if a substantial part of these pop songs were actually hip hip songs (e.g. “Fuck Me, Fuck You” by Blackstreet and “Money Good” by Chester Bennington). This means that the faulty classifications are partly due to a deficiency in our “objective” reference point rather than a deficiency in the classifier.

The high accuracies for all three models indicate that it makes sense to regard hip hop lyrics as a taxonomic class. Disregarding cognitive and cultural context turns out to be a fairly good premise when examining hip hop lyrics. To what extent this applies is an open question and does not appear from this study. But the results set a baseline from which to further explore hip hop as a taxonomic class, a cognitive category, and a cultural unit. Additionally, the selected low-level features were shown to be more indicative than the general topics found in the lyrics. This indicates that the conventions of hip hop as a lyric genre are not wholly driven by meaningful properties, but are actually to a larger extent structural features with no inherent meaning. Whether this is a general trend in music lyrics is a subject for further exploration. It is very likely that some genres of music are so diverse that it does not really make sense to consider them taxonomic classes. The method of inquiry used in this study could also be expanded to other written domains, i.e. non-literary prose, movie scripts, news articles etc., and the questions about genres, meaning, and automatic genre classification raised in this paper are just as relevant for auditory mediums such as the actual sound components of music as well as visual mediums like movies and paintings.

Reflections on methodology

Meaning and division of features (L, A)

The interpretation of the results as indicative of non-meaningful features being more indicative to hip hop than meaningful features rests on four assumptions: that our definition of meaning is adequate, that high-level features are more meaningful than low-level features, that we have correctly operationalized these features, and that we have exhausted these categories. While problems concerning correct operationalization especially applies to high-level features, exhaustion of the category is a problem that is particularly relevant to low-level features. One can think of several ways of operationalizing high-level features such as topics and sentiment while low-level features are often unalterable. Especially features such as 'Number of words' are easily quantifiable, and while features like 'Number of choruses' and 'Number of lines' might be more controversial, they are not as open to interpretation as high-level features. But one could think of many more features that fit our definition of low-level. One could also make a case that style features such as rhyming patterns and part-of-speech features such as word classes and swear words, but an equally persuasive case could be made that these kinds of features are high-level. Perhaps the division of lyric features into levels of complexity proposed in this study ought to include intermediate levels. Certainly, a more consistent and robust theory of meaning in relation to lyric text would help with substantiating this distinction.

Another limitation of this study is exclusively focusing on the symbolic representations of music and not its audio qualities. Although a limited scope allows us to draw more exact conclusions for hip hop lyrics as a particular aspect of hip hop music, it decreases the validity of expanding these findings to hip hop as a whole. But lyrics have the advantage of being generally more data sparse, available, and codeable when compared to auditory aspects of music. As automatic audio information retrieval becomes increasingly viable, we would expect the implementation of audio features in multi-modal models to outperform purely lyrics based models.

Classification troubles (L, A)

A big turning point in any classification study is the chosen golden standard for the classes. This study used the list of Wikipedia articles classified as hip hop artists and assumed all their songs could be called hip hop. Wikipedia's classifications are interesting as it is a compromise of many different people's classification. We anticipated this approach would give a fair assessment of genres as it relies on many different opinions instead of a few. Also, the editing occurs naturally according to recent changes and new cultural movements. Despite that, there are definitely problems with this approach, artists sometimes change their style or write songs outside their typical genre. By excluding the hip hop genre to artists identified as hip hop artists, the models consistently scored lower on identifying non-hip hop correctly, because our golden standard was wrong in some cases. Many pop songs also feature a rapper, and many rappers also produce traditional pop or R&B songs. All these nuances are regrettably lost and will bring noise into the model. Also, we gave hip hop songs priority. That is, we first collected all the hip hop songs available, and then the other genres. Then when removing all duplicates the one collected first was kept.

For the strategy of using Wikipedia articles of artists, some possibilities for improvement are available. Genius.com has a tagging system for songs, which could be used to check whether the users of Genius.com would classify the song as hip hop. Another idea is to check whether all the artists involved in the song is classified as hip hop artists, this would catch the mixed genre songs. However, there will always be some uncertainty and misclassification as genres of course are controversial.

Messy data (L, A)

This study relied on crowdsourced lyrics from the website Genius.com as data. Therefore, the data is not optimised for machine processing. Genius.com does not offer professionally written lyrics, rather it is written by fans that try to imitate the style and voice of the singer. Hence our data included many different words not normally seen in the English language. Especially onomatopes and slang words come in very different forms (e.g. 'aah', 'ahh', 'ahhh' or 'yall' and 'wattup'). Even though these might be similar, the model accounts them as different, and we can not control that these are written in the same way across different songs. Furthermore, expressions like 'repeated till end' and '2x' are common in written lyrics, which of course sends the wrong information into the

model. These repetitions could potentially be interesting, but are in many cases unavailable without heavy preprocessing. Similarly, the number of lines in the text are sometimes written in unexpected ways. For example, if a song have particularly long pauses between lines, they will be represented by extra empty lines, which will count as another full line in the model. This results in the songs with least words per line are actually just songs with long pauses between the lines, which is comparable but not interchangeable.

A considerable worry for this study is that we only measure differences in writing lyrics. Even though we have pulled all lyrics from the same site, which presumably have a core demographic, it is highly probable that fans of different genres write lyrics in different ways to distinguish their genre from others. As mentioned earlier, we observed a distinct way of writing present participle in hip hop songs (without a 'g' in the end, seeing/seein'). This is just an example of a general trend to modify the spellings of words in hip hop lyrics including 'ya' instead for 'you' or 'your' and 'cuz' instead of 'because'. It could be that a model like LDA would find common verbs written in this style very indicative of the genre and therefore rely on *the way it is written* instead of the *actual words*. Whether or not this difference in style actually captures a worthwhile difference in terms of the semantics is debatable, but it would be preferred to have a standardized way of writing these verbs, as it results in substantial noise.

Assessing the LDA-model (L, A)

With an accuracy of 85.85%, the LDA-model was our lowest performing model. This is however not surprising as it was trained on only 10% of the total dataset. With the small training set in mind 85.85% accuracy is still a respectable score. Are there any ways of increasing the accuracy of the high-level model? Obvious ways include cleaning the dataset more thoroughly and increasing the size of the training set. A more interesting choice was our number of topics assumed in the corpus, often referred to as *k*. In this study, we chose a model with only 2 topics, presumably hip hop and non-hip hop. Models with more topics were tested, but never with better accuracy than the simplest. We are of course only interested in binary classification in this study, so it made sense to use two topics from the beginning. However, one could imagine a model with 100 topics representing many different subgenres of songs in the data (like the model shown in Blei, 2012). This model is more complex and only works by assuming the genre-blend in the dataset is relatively

low since each topic must be put in either hip hop or not. A big problem with this model is the complexity it presents to the developer. It has to be a human who identifies each topic, which would require deep knowledge of the music and subgenres present in the dataset. This is the tradeoff for having a more complex model that is able to catch many different subgenres: a human has to identify and know these subgenres. Realistically, it would require a highly customized set of stop words (the words the model should not consider for labeling), which can only be done by a lot of testing, which is time-consuming for a model this complex. In conclusion, the simplest model worked sufficiently well and was easy to manage. A more sensitive model would require a solid understanding of the dataset and more time optimizing, but could potentially yield higher accuracy.

The naivety of sentiment analysis (L, A)

The central assumption of sentiment analysis, at least the naive version used in this study, is that a text's sentiment can be assessed by accumulating sentiment scores for each word in the text. There are several general problems with this assumption. First of all, sentence structure and morphology are ignored which means most importantly that the sentiment analysis misses a lot of what makes language meaningful (e.g. negations and differentiation between subject and object are not taken into account). Secondly, context is largely ignored in sentiment analysis. Using a generic, general-purpose dictionary assumes that sentiment scores ascribed to different words are the same for all people in all contexts. It is evident, though, that no word has a universal meaning, nor a universally ascribed sentiment. Especially for hip hop music in which the language is often hip hop nation language (HHNL) or heavily inspired by HHNL (Alim, 2004), words might express very different sentiments. Words that normally are regarded as negative can have a positive meaning in hip hop (e.g. 'gun' and 'niggers'), and many words will not be attributed a sentiment score either because they are not found in the sentiment dictionary or because a word is a spelling modification of a word contained in the sentiment dictionary. This is a particular problem for hip hop because the lyrics often use alternative spelling (Alim, 2004). The alternative spelling and spelling of HHNL also creates problems when stemming words. We would expect that a corpus-specific sentiment lexicon would increase the performance of sentiment analyses on hip hip lyrics significantly.

Oudenne and Chasins (2010) identified a problem particular to song lyrics: that they often express conflicting emotions with one emotion being prevalent in the majority of the song, but then resolve

into the opposite emotion by the end of the song which skews the sentiment score of the song. Though the effect is expected not be a big influence in hip hop in an of itself, it is problematic when taking other genres into account. Irony is another example of sentiment analysis failing to identify the intention in a song. Any irony, sarcasm, or mockery is taken at face value. This is also true for metaphors and intertextual references. Because they depend largely on the sociocultural context of the song, these problems do not seem likely to be resolved in the immediate future.

Conclusion (L, A)

This paper sought to illuminate the possibilities and obstacles of classifying songs based on linguistic characteristics. If hip hop as a lyric genre were indeed to be considered a taxonomic class, we should be able to accurately predict via machine learning whether a song was a hip hop song or not. Using song lists from Wikipedia, 105,000 song lyrics obtained from Genius.com, topic modelling, and sentiment analysis, three predictive models were constructed via logistic regression: A high-level model using topic affiliations alone as a feature since sentiment scores were found to be insignificant, a low-level model using six text-statistic features, and a combination of the two mentioned models. The accuracy of the models' predictions were 85.85%, 87.67%, and 90.29% respectively. This indicates that it makes sense to regard hip hop lyrics as a taxonomic class and that the distinctiveness of hip hop lyrics, when compared to other genres, is due to non-meaningful features rather than meaningful features.

The paper's main contribution is showing that modern computational methods can be utilized when trying to answer overall questions about genres as well as questions about individual genres' idiosyncratic properties. There are some serious methodological hurdles to be solved before the suggested approaches become viable. First of all, defining meaning and feature levels to substantiate data analyses proves difficult and currently rely on assumptions whose validity is open to discussion. Secondly, improving the high-level features, like topic and sentiment analysis, and making sure that low-level features are exhausted is important because small adjustments can be decisive factors. Furthermore, tagging systems must improve as they are the semi-objective reference points to which a model can be compared. Future studies can be carried out to investigate the limits of hip hop lyrics' taxonomic nature, to adapt the analysis method to other genres and

other mediums, and to explore how alterations in high-level and low-level features affect performance.

References

- Alim, H. (2004). Hip Hop Nation Language. In E. Finegan & J. Rickford (Eds.), Language in the USA: *Themes for the Twenty-first Century* (pp. 387-409). Cambridge, UK: Cambridge University Press.
- Anbari, S. A. (2010). *The Genres, Prosody and Pragmatics of Rap.* Munich, Deutschland: GRIN Verlag.
- Attridge, D. (1995). *Poetic rhythm: an introduction*. Cambridge, UK: Cambridge University Press.
- Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL (1) (pp. 238-247).
- Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent dirichlet allocation. In Advances in neural information processing systems(pp. 601-608).
- Edwards, P. (2009). *How to rap*. Chicago, Illinois: Chicago Review Press.
- Feuerriegel, S. & Proellochs, N. (2013). *Dictionary-Based Sentiment Analysis*. R package, version 1.3-0.
- Halliday, P. M. (2017, September). *A Theory of the Musical Genre: The Three-Phase Cycle*. Paper presented at the Proceedings of the 10th International Conference of Students of Systematic Musicology (SysMus17), London, UK. Retrieved from https://sysmus17.qmul.ac.uk/wp-content/uploads/2017/08/mendoza_halliday_three_phase_cycle.pdf
- Hanna M. Wallach, Topic modeling: beyond bag-of-words, Proceedings of the 23rd international conference on Machine learning, p.977-984, June 25-29, 2006, Pittsburgh, Pennsylvania
- Liang, D., Gu, H., & O'Connor, B. (2011). *Music genre classification with the million song dataset*. Machine Learning Department, CMU. Retrieved from http://www.ee.columbia.edu/~dliang/files/FINAL.pdf

- Mayer, R., Neumayer, R., & Rauber, A. (2008, September). Rhyme and Style Features for Musical Genre Classification by Song Lyrics. In *Ismir* (pp. 337-342).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).
- Olivo, W. (2001). Phat lines: Spelling conventions in rap music. Written Language & Literacy, 4(1), 67-85.
- Oudenne, A.M. & Chasins, S. E. (2010). *Identifying the Emotional Polarity of Song Lyrics through Natural Language Processing*. Retrieved from https://pdfs.semanticscholar.org/22c6/b9e01e33a779c922aea3af32f6807127522f.pdf?_ga=2.88411803.1686587613.1514500812-1465123820.1514500812
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825-2830.
- RStudio Team (2015). RStudio: Integrated Development for R. Rstudio, Inc., Boston, MA. URL: http://www.rstudio.com
- Sahlgren, M. (2008). The distributional hypothesis. Italian Journal of Disability Studies, 20, 33-53.
- Silver, D., Lee, M., & Childress, C. C. (2016). Genre Complexes in Popular Music. PLoS ONE, 11(5), e0155471. http://doi.org/10.1371/journal.pone.0155471
- Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.
- Volcani, Yanon; & Fogel, David B., "System and method for determining and controlling the impact of text", published June 28, 2001.
- Yager, R. R., & Zadeh, L. A. (Eds.). (2012). *An introduction to fuzzy logic applications in intelligent systems* (Vol. 165). Springer Science & Business Media.

Appendix

Appendix A: Wikipedia articles for hip hop artists

Articles are retrieved November 28, 2017

- List of hip hop artists (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/List_of_hip_hop_musicians
- List of hip hop groups (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/List_of_hip_hop_groups

Appendix B: Wikipedia articles for pop artists

Articles are retrieved December 5, 2017

- Category: American pop artists (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/Category:American_pop_singers
- Category: British pop singers (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/Category:British_pop_singers
- Category: Canadian female pop singers (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/Category:Canadian_female_pop_singers
- Category: Canadian pop singers (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/Category:Canadian_pop_singers
- Category: English pop singers (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/Category:English_pop_singers

Appendix C: Wikipedia article for alternative rock artists

Articles are retrieved December 6, 2017

- List of alternative rock artists (n.d.). In *Wikipedia*, from https://en.wikipedia.org/wiki/List_of_alternative_rock_artists