ERASMUS SCHOOL OF ECONOMICS

DEPARTMENT OF BUSINESS ECONOMICS - MARKETING

MASTER THESIS

It's All in the Lyrics:

The Predictive Power of Lyrics

Date : August, 2014

Name : Maarten de Vos

Student ID : 331082

Supervisor : Dr. F. Deutzmann

Abstract:

This research focuses on whether it is possible to determine popularity of a specific song based on its lyrics. The songs which were considered were in the US Billboard Top 40 Singles Chart between 2006 and 2011. From this point forward, k-means clustering in RapidMiner has been used to determine the content that was present in the songs. Furthermore, data from Last.FM was used to determine the popularity of a song. This resulted in the conclusion that up to a certain extent, it is possible to determine popularity of songs. This is mostly applicable to the love theme, but only in the specific niches of it.

Keywords:

Content analysis, RapidMiner, Lyric analysis, Last.FM, K-means clustering, US Billboard Top 40 Singles Chart, Popularity

Table of Contents

1. Introduction

Music is considered as one of the biggest parts of our daily lifes. One can come across it in a wide variety of situations, e.g. in the car, in the mall or at home. Music, however, is also one of the most competitive industries. There are a lot of home artists trying to penetrate the industry while the big artists keep the charts filled.

One of the main issues these days is that the total music industry seems to be in a decline. A recent news story in the New York Post written by Atkinson (2014) indicates that global music sales have decreased by 4% in 2013 alone. However, there are positive signs in the article. Based on the news article, it becomes clear that this is not the same on every continent. Declines can, for example, be seen in Japan while in the USA and Europe, growth is apparent. This could possibly be explained by better advertisement or more opportunities for success, such as streaming services.

Through the services of Spotify, iTunes and YouTube, one can easily become a self publishing artist, while labels have more distribution channels. It is therefore more important than ever to find out which songs are likely to succeed and which are more likely to fail.

This research will therefore try to find out, *which themes are more likely to become popular songs based on its lyrics.* It is important to recognize that popularity does not necessarily result in profits. The goal of the research study is to provide a method to determine which of the themes people are currently more interested in.

It was found that there are certain correlations between the lyrics and the popularity of a song. It can also be argued that specific niches within the love themes are able to capture a bigger audience compared to their counterparts. It does not matter whether these counterparts are love related or not.

2. Motivation

This chapter will focus on the motivation of this research. The will be done in two parts, first the academic contribution will be discussed, after which the managerial contribution will follow.

2.1. Academic contribution

This research is important for academic research since it shows another place were text analysis could be useful. At the moment, text analysis is being used in a very limited environment, such as looking at products and conversion rates (Ludwig et al., 2013) or product positioning (Lee & Bradlow, 2011). Due to the computerization in the last decade, text analysis became a lot easier for the researchers and this can be applied in a much broader setting.

An example of the limited research available about lyric analysis is the research of Knobloch-Westerwick et al. (2008). They looked into whether rebellious categories, mostly based on hostility and aggressive, were more common in the lyrics they analyzed. Since the lyrics that they analyzed were all part of some top chart, it would mean that rebellious related lyrics would be the norm and therefore more successful in the two genres that were considered, Rap/Hip-Hop and Rock. Since such a relationship was found, it could be argued that certain categories of lyrics could be more successful in other music genres.

This research will focus on automated text analysis of lyrics. One of the main academic contributions for this research is that there is almost no research available. As mentioned before there are some other researchers who have used similar methodology, but it has not often or never been applied to lyrics before.

Another issue that has been present in the past and is still partially an issue at the moment is the difficulty of acquiring data to perform this kind of research. There are a lot of different ways to measure success and most of the data is not publicly available. Researchers therefore have to choose to use suboptimal options or not to do the research at all. In the past 10 years, the internet have made it a lot easier to acquire certain data, such as the lyrics themselves. The research opportunities have therefore increased a lot since the last few years and this is the reason why the timing of this research is relatively good. In the method sections it will be discussed how these opportunities are going to be exploited.

2.2. Managerial contribution

The managerial implications of lyric analysis can be quite huge, but some caution is advised because it is a relatively new area of research. However, it may provide some methodology for future research and companies.

Assuming a model is found, it would provide insight on which direction future research should focus. Furthermore, the model might be able to compare successful themes to less successful ones. Artists and label managers would be able to use this information to adjust their strategy. Besides the larger stakeholders, there is another group who could benefit from this research. Independent musicians can adjust their strategy just as much, because of the low cost of this research . If they are looking to be the next top artist, they may want to focus on a different theme than what they are currently working with.

The ultimate goal of this research is not to create a model that will predict the next hit song, based on specific themes. The goal of this research is to create a method of acquiring data about the current theme preferences of consumers. This research is therefore meant to see whether a proof of concept can be created based on clustering. It is assumed that people do not immediately shift their preference, however this assumption cannot be guaranteed in the long run. The results regarding which themes are most important are therefore less relevant, but it can still be of interest for the short run.

This research is looking more into the possibilities of applying a method that can potentially become very relevant for future marketeers. The case study created by Elberse et al. (2005) shows the interesting concept of how to combine music and math to better predict the success rate of the songs. The case implies that the cost of introducing a new song to the public is an expensive endeavour with a relatively low chance of a high return. This methodology could therefore be used as a research or marketing tool by record labels. Knowing what your customers want is an issue that marketeers face every day and this research is trying to provide an additional means to this end.

Another method for text analysis came from Wathieu and Friendman (2005). They wrote a case study about sentiment analysis and how it can be useful to classify product reviews. In the case study, the company Intelliseek provided a sentiment analysis of texts of other companies found on the internet. However, it is only used in a limited amount of industries. This case study argue that text analysis does not work in all industries but when it provides benefits, it can be very beneficial.

The most important lesson to be learned here is that the industry needs to be willing to adapt to new marketing tools. Therefore, this research is not trying to replace the current marketing tools. Instead, if a viable method and/or model is found, the goal is to support current marketing tools.

3. Literature review

Since research in the area of song lyrics is rather limited, this research will look into two different fields for the literature review. The first focuses on how text analysis has been applied in the past. This way some insight can be gained on how text analysis can be used to analyze lyrics and what kind of results can be expected. The other part focuses more on what is currently known in the research of music popularity and what it is driven by.

3.1. General text analysis

Ludwig et al. (2013) performed some research in the area of how customer reviews change the conversion rate of books on Amazon.com. Although the research is not directly related, they did confirm earlier studies that more negative reviews cause a more consistent change in conversion rates compared to positive reviews. Logically this could suggest is that people dislike more negative subjects and therefore it can be related to this research. If this would be the case, then results would be found where love, in a more negative context such as break-ups for example, would sell less than the positive songs about love. They also acknowledge that conversion rates are affected by the way a review is written. When this happens in a similar linguistic style compared to your own, the conversion rates seem to be positively affected. This while the opposite also holds true when the linguistic styles are different.

The research of Lee and Bradlow (2011) confirms that text mining and analysis can be a valuable tool for confirming ideas, positioning, and concepts. Companies can use customer reviews to confirm the results of other tools. It also allows for confirmation that certain strategies might work or not. They do however recognize something in their limitations section which this research needs to keep in mind. They talk about the concept that context is important for understanding a lyric or a sentence. Therefore, when considering the output of any text analysis software, it is necessary to consider this with great care.

Lee and Bradlow (2011) used k-means clustering, which will be discussed in detail in the methods section, in their research of product reviews. They used this because it is well known within the marketing community and it is easy to identify the clusters of similar product reviews.

Another example of a study which uses text analysis was done by Constance and Silverblatt (2014). Their study was about whether sustainability is being taught in colleges. The most interesting part that is applicable to this study is that they argued about how important context can be when analyzing texts. When considering their research method and design, they explicitly chose to do the text analysis by hand. The method was, however, relatively simple since they used a word count of specific words in a dictionary to determine how much sustainability was discussed. This could be a relatively simple alternative approach for future research.

The research by Constance and Silverblatt (2014) does give some insight in the importance of context. Therefore using only a word count will not be appropriate. The fact that context matters therefore got some more value and caution is advised when interpreting (strings of) words.

3.2. Music and lyrics

One of the major questions is whether lyrics are actually heard and processed. If this would not be the case, the whole concept of analyzing lyrics to predict success would be redundant. Hansen and Hansen (1991) described this in more detail by arguing that young adolescents might not completely understand and process the lyrics but do understand the basic gist of the song within the metal genre when listening to it. Hall (1998) supported these findings and adds the comparison of recall versus recognition of rap and metal lyrics. Some evidence contrary to these findings can also be found, e.g. Greenfield et al. (1987). Here, it was argued that sometimes the general knowledge of (early) adolescents might not be enough to comprehend the lyrics correctly; thus resulting in a misunderstanding of the lyrics presented.

Johnstone and Katz (1957) wrote in their research about how music taste differs depending on personal circumstances. The test group consisted of teenage girls. Items that were included in their analysis are the neighbourhood where they lived in, popularity and also the music taste of their friends. The research showed that music is highly dependent on the environment where people reside. The neighbourhood norm defines what music is popular and the popular girls tend to adjust their music choice to the norm. Groups of friends also shared a similar vision in music and DJ's. The researchers however could not confirm whether this evolved over time or whether these persons got closer to each other because of their similar music taste. These results might be applicable on a larger scale. Thus, society as a whole might pressure each other with certain music preferences. It would be one possible explanation for why certain themes might be more common than others in popular music.

Lewis (1993) published a book with a multitude of paper style chapters in it. In this book, a chapter was written by Charles Jaret (p. 174-185) in which he described how country music popularity is affected by a variety of variables. Although there are some limitations to the research, and they acknowledged that the theme of the song on its own is not sufficient to explain success, they do find some important themes for female singers. For a female singer, they for example find that sad love, rambling theme and sexual themes have similar effects on the number of weeks present in a hit list and the highest position acquired on a chart. A rambling theme was defined as being corrupted, working, having a good time, driving and travelling. Violence, make up, break up and music themes had no significant effect.

Carey (1969) also dedicated some time to conduct research on how lyrics change over time based on the current standards of society. He suggested that lyrics and themes are not stable in the long run. One of his interesting findings was more focused on advertisement budgets. He said that songs with a higher advertisement budget are more likely to succeed.

Hobbs and Gallup (2011) were able to identify a relationship with one of the themes which was partially considered in this research. They researched whether reproductive messages, themes about body parts, love, sex and so forth, are significantly present in the Billboard Top 100 charts of pop, country and R&B. They also examined whether songs with reproductive messages are more likely to end up in the top 10 or not. Both of these hypotheses were confirmed in all three genres.

Another paper which did some basic text analysis was written by Freudiger and Almquist (1978). This research was interested in whether genders were portrayed differently in lyrics. The text analysis in this paper was done by hand and may be biased but it may be able to capture the context-specific words and settings better. The paper did however find some differences in the data set in how popular songs were portraying males and females: "First women are presented much more positively than men, and, second, there is more variability in feminine traits across genres than there is in male traits." (Freudiger and Almquist, 1978, p. 63) . This research is therefore important to keep in mind when determining the themes in the lyrics that will be analyzed. This is due to the fact that the gender of the subject in a lyric can play a significant role in how lyrics are perceived.

3.3. Theoretical framework

The theoretical framework is mostly used to support the hypothesis that lyrics contribute towards the popularity of a song. Therefore, the main theoretical framework that will be used in this research was written by Schwarz and Clore (2007). Their paper divided the feelings and thinking part of their theory in three approaches: "The first approach emphasizes the experiential quality of feelings and addresses their informational functions. A second approach emphasizes the thoughts that accompany feelings, whereas a third approach emphasizes hard-wired processes, focusing on the somatic components of affective states." (Schwarz & Clore, 2007, p. 385)

The major interest for this thesis lies in the first two approaches. The first approach deals with the differences between, for example, emotions and feelings. Since emotions are of relatively short durations and can be triggered by impulses, this may explain why a certain song is liked or disliked. People have a certain feeling that can trigger the emotions that come with the song. This might result in people liking the song more or less. Part of the song is related to the lyrics since people usually have a general understanding of the lyrics (Hall, 1998; Hansen & Hansen, 1991).

Moods, cognitive experience and processing fluency are also part of the first approach. This, however, is much more unlikely to affect the likeability of a song in the short run. For some of these things to happen, the lyrics and/or song must be known. It requires more of a thought process compared to emotions. Moods can be caused by songs within the same theme. These concepts are dependent on the longer term vision and how people judge a song. This includes not only hearing the song repeatedly but also hearing it within a certain playlist.

Another pitfall is that people might attribute feelings towards a certain song while those feelings are not caused by the song, but by a third party instead. Unfortunately, this cannot easily be controlled. However, through randomization of the listeners this should be relatively limited where the positive and negative feelings towards a song should be able to cancel each other out.

The second approach is mainly focused on how feelings affect the kind of thoughts that come to mind. During this literature review, the basic conclusion is made that thoughts are not as 'easy' to model without making a significant amount of assumptions. Therefore, this has resulted in an unworkable model. This shows how difficult it is to predict any of the results.

Another part of the paper by Schwarz and Clore (2007) discussed how feelings affect judgements. The most important part of this section is that people in general think that the feelings they have are directly related to the task or impulse at hand. One of the few exceptions is when people are able to attribute their feelings to something else; thus, they need input on what it could be otherwise. Feelings are important when the decision has to do with the preferences of the person.

How feelings can affect judgements may have a significant effect on the repeated listening to a song. By attributing specific feelings at a point in time to a specific song, genre, theme or otherwise important attribute of a song, the listeners can be more or less likely to listen to a specific song again. Therefore, the whole effect of emotions and feelings may cause different kinds of effects on the listeners. Emotions might determine whether people finish the song they are currently listening to, while feelings might influence people to listen to songs for a longer period.

4. Data - Sources

4.1. Introduction

This research will focus on how the popularity of a pop song can be predicted by analyzing the lyrics. Since it is not possible for this research to acquire the sales data of specific songs, another measurement had to be chosen. Popularity is usually related to how often people listen to the songs while success is linked to monetary income from sales. There is still however an existing definition problem. The definition of popularity is complicated since it nowadays consists of a multitude of factors. Therefore, the first discussion has to focus on all possible ways to (directly) acquire or interact with music. It is assumed that people who engage in these activities do really choose to do so like; it is considered a conscious decision. Therefore, walking into a store with music on is not considered a conscious interaction with music.

First, there is the radio and music television, that can more or less choose between a wide range of singles to broadcast. Therefore, the possible set of songs being played is highly limited to the current DJ or producer and only singles can be played. Many radio or TV stations limit themselves in the genres of music they provide to their listeners and adjust their playlists based on that. They will only use sources which come close to their target group.

Another option to consider is the sales data of CDs, DVDs, concert tickets, downloads and so forth. In general, this would give a fairly accurate overview of what is considered more popular. This would have been a very good measurement for success; however, it cannot always be specifically attributed to the popularity of a song. The popularity may be due to its artist or the album. Recently, it has also been shown that some of these measurements are being manipulated by the artists[1]. Therefore, some extra consideration is needed if a measurement is used to determine whether the popularity actually reflects on the artist.

Another very recent phenomenon is the use of streaming services such as Spotify and YouTube. Instead of buying a specific song, these streaming services allow listening to a broad catalogue of music albums and singles. The only limitation of these kind of services is that they might not offer all the music people want to listen to because they cannot get a license for every song due to

[1] Dutch consumer programme “Rambam”, 17 February 2014.

financial constraints of those companies. The streaming keep track of how often people have listened to certain tracks for payment purposes. Some of the streaming services show these numbers publicly.

Finally, there are the more questionable ways to acquire music. This often results in downloads or streams that have not been paid for in any way. These ways of acquiring music are very difficult to track for record labels. Therefore, they do not show up in any official data sources.

Now that all the options for acquiring music are considered, the remaining question is what is considered as popularity of music. Basically, it is all of the above combined, but this is difficult to measure. Many of the options above are already difficult to measure on their own, let alone when they are combined. Therefore, this research will try to combine as many of them in one dependent variable as possible.

4.2. Dependent variable

As explained in the previous section, one of the hardest features of the dependent variable is measurability. There are a lot of opportunities, but not all are as viable. One issue could be for example, that the data of the dependent variable is not acquirable. The latter holds true, for example, for the specific numbers of downloads of streams.

There are a few possible ways to circumvent this. One is to examine how the hit lists are set up and used as measurement of popularity. The variable however will then become ordinal since all hit charts measure rankings based on sales in units or number of streams compared to each other. Similar problems also exist when trying to acquire data from music television or radio since they mostly publish the rankings of most played songs. When using any ranking the focus of the research may be quite limited as it only shows the top of the notch at a specific moment. This can partially be solved by making it more period based by examining multiple weeks or years of rankings.

However, the use of an ordinal measurement is not preferred by this research. Due to the internet, some new websites have emerged and one of them is called "Last.FM". Last.FM allows users from all over the world to upload ("scrobble") the music they are currently listening to on the website. When they listen to a specific song for more than 30 seconds, it is added to their profile as listened to. The software which is used works with almost any operating system and music player. In some players, it is even natively build in, such as in Spotify. Finally, a lot of mobile devices and mobile music players can also update the Last.FM profiles. Last.FM allows developers to access the play count of every song they have in their database through the Application Programmable Interface ("API"). This way, you can access the total amount of times that a song has been played. By using Last.FM, most of the issues that other measurements have can be minimized while keeping any new ones to a minimum.

The benefit of using this approach is that the raw data will actually be an integer. It gives the exact differences between the two data points. It also allows a larger set of music lyrics to be analyzed since it does not have to be collected by hand. Finally, the play count is based on everybody who uses it, wherever the song originated from, including programmes such as Spotify and iTunes. The only requirement is that when it is from an illegal source, that the meta data is correct. This is considered a fair assumption because otherwise there still would be no way to measure these listeners. Furthermore, people regularly correct the meta data themselves so they know which song they are listening to. The downside of using Last.FM, however, is the partial bias because of the users contributing towards the play count since it does require specific software to be installed. In summary, the advantages outweigh the disadvantages and therefore this approach is considered.

4.3. Independent variable(s)

The independent variable will be created based on the lyrics of the song in question. The independent variable will be specified as specific themes which will be determined by using text analysis software, such as RapidMiner (Feldman & Sanger, 2006). RapidMiner can cluster various songs together by means of k-means clustering. These can then be used as independent variables.

4.4. Control variables

Five possible control variables have been identified. The first is the length of the song. This will probably not be a linear effect, but there is a possibility that longer songs are appreciated more compared to shorter ones. This thought was based on some initial pilot regressions. Longer songs might be considered more meaningful or deeper. There is, however, a chance that this control variable will be insignificant. Due to the absence of a linear effect, this control variable is transformed to a natural logarithm. Through some initial pilot analysis, the best fit has been determined. The control variable is based on the Last.FM API, since they also show the duration along with their play count.

The second control variable has to do with the release date. It is fully understandable that a song which has been published 10 years ago will have a higher probability of a higher play count compared to the songs which were released only 2 months ago. Therefore, the natural logarithm of number of days since its release up to May 1st, 2014 was considered to be more appropriate. This control variable is based on the Last.FM API since they also show this information along with their play count.

The third control variable is the genre of the song. The genre can have quite a significant effect on whether people listen to the song. Although the genre analysed will mainly be pop music, a song can be part of multiple genres. There is also a chance that certain topics are more common in certain genres and there is a chance of multicollinearity. Therefore, a principal component analysis between the themes (independent variables) and the genres has been done.

The fourth control variable focused on whether the vocalist is male or female. It was argued by Lewis (1993) that the voice of a song can significantly influence the sales or popularity of a song in certain genres and themes.

Finally, it could also be the case that specific artists show a significant effect on popularity. This will be measured by means of the total listeners of an artist. Last.FM provides this information through their API, which includes every listener that listened to any of the songs of that particular artist. Therefore, if an artist has a broader set of popular music, this will probably not be correlated as much with the song listener total as compared to when the artist has only single US Billboard Top 40 Singles Chart song. When the latter is the case, there is a higher chance that the majority of the listeners has just heard one song of the artist while this is not necessarily the case with an artist with multiple hits. Finally, just as before, the natural logarithm will be considered for this control variable as well.

4.5. Lyric selection

This thesis focuses its efforts on the top 40 chart music. Music lyrics were selected based upon multiple US Billboard Top 40 Singles Charts between 2006 and 2011. The five year limit was chosen due to the changing preferences of people as described by Carey (1969). The sample size is around 900 lyrics. Using the US Billboard Top 40 Singles Chart provides an assurance that the music present in the subset have or have had some mass attention. This is important since the Last.FM play count will be used as a dependent variable. Some groups, such as elderly, might be underrepresented on Last.FM and may therefore cause a bias. Using a top 40 chart will hopefully reduce this bias, because it is a representation of the population.

5. Research method

The research method is based on three different aspects. The first step is to collect the relevant data. This consists of two parts: the song lyric Last.FM data collection. Both will be acquired by means of a custom written web crawler.

The second part of the research consists of using text analysis software. For the text analysis, RapidMiner has been used for K-means clustering. RapidMiner provides this research with the opportunity to analyze a large batch of text in a relatively short amount of time. Besides this, RapidMiner is available for free and can be used to replicate similar research. The results which RapidMiner provided are also in a format which can be used in other software, such as Microsoft Excel and SPSS.

The final part of this research consists of checking the assumptions of the linear regression and analysing the results. Diagram 1 gives a summary of the different parts.

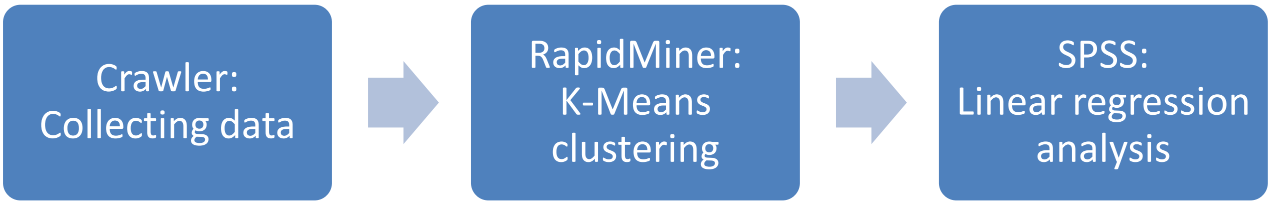


Diagram 1: Research method: Process

5.1. Tools

Due to the wide variety of versions that are available for both software packages and possible differences this can cause, this has to be noted. RapidMiner version 6.0.002 has been used for the purpose of analyzing the various lyrics. Since RapidMiner does not support text analysis by default, a plug-in is needed. The plug-in used is the Text Mining Extension 5.3.2. For the regression analysis, SPSS 20.0 has been used.

5.2. Crawler

The crawler is custom-written in Hypertext Preprocessor ("PHP") with the use of MySQL as database software. PHP is a web programming language which allows access to the source code of various websites, including the ones where lyrics can be acquired. Furthermore, it also has the capabilities of accessing the Last.FM API. Both data sources can then be stored in a MySQL database which can easily be downloaded into an Excel file and can then be used in RapidMiner and/or SPSS.

5.3. RapidMiner

The analysis with RapidMiner has been split up in two parts. The first focuses on preparing the data while the latter is the K-means clustering.

The data preparation is based on the book by Hofmann and Klinkenberg (2013). The process followed the following steps: First, the text was prepared before it was processed. The function, "Process Documents from Data" allowed this. In the data preparation, every song was completely tokenized based on non-letter characters. This resulted in an array of words for each lyric which was analyzed. Stemming was the next step, in which similar words were adjusted by RapidMiner, so they can be grouped appropriately. To make them more comparable, all letters were transformed to lower cases. This step was followed by the removal of any irrelevant words. RapidMiner has an internal dictionary for stop words, which was used. These were all deleted. The next step was to delete two words which were deemed non-relevant for the content of the lyrics: "chorus" and "verse". Finally, any word consisting of only one character was deleted as well, since it was unlikely that they will add anything of value to the analysis without the context. The total data preparation as shown by RapidMiner can be seen in Image 1.

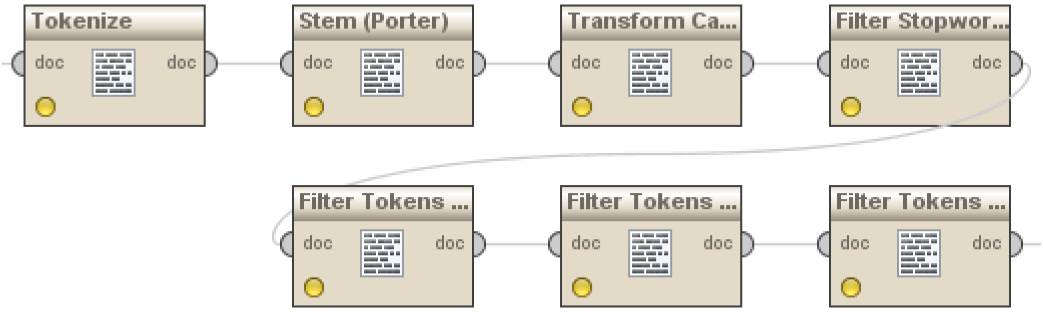


Image 1: Data preparation

The word count happened based on the TF-IDF[2] algorithm. TF-IDF[2] looks at the data in two steps (Ramos, 2003). The first is the term frequency, which is the relative frequency of a word in one lyric. The second part consists of taking the natural logarithm of the inverse document frequency. It looked at how often a specific word exists in all the lyrics. TD-IDF was ran with an absolute prune method with a minimum of 2 and a maximum of 9999.

The next step was to perform the K-means clustering. Clustering is necessary to make the lyrics comparable in any quantitative analysis. These similar lyrics can then be analysed together in, for example, SPSS. When the clustering in RapidMiner was completed, it provided a result with a set of songs that belong to a specific cluster. Due to the method it is impossible to have a single song in multiple clusters at once. The whole process, as shown in RapidMiner, can be seen in Image 2.

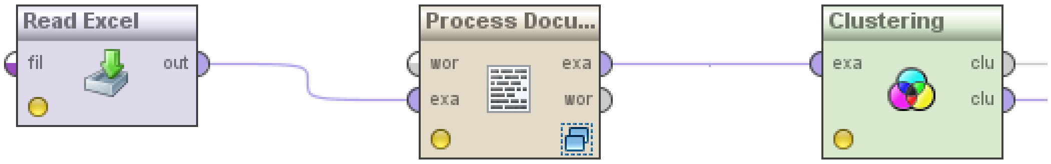


Image 2: Clustering

With the cluster information, it is possible to create a dummy variable for every existing cluster. This data was then imported into SPSS. The main issue here is that the number of clusters must be determined by the user of RapidMiner. To determine the number of clusters appropriately, this research study took into consideration the adjusted R-square. The adjusted R-square determines the explanatory power of the model based on the number of variables in the model and how much they benefit the model, which is presented in a percentage.

Every time the analysis was run in RapidMiner, all the information regressed in SPSS to determine the explanatory power. Depending on whether the explanatory power has improved or not will determine whether the number of clusters will increase or decrease. This way, the highest possible explanatory power was acquired while also keeping a consistent rule in mind when determining the number of clusters.

[2] The exact formula is (Ramos, 2003): (Number of times word x is present in this lyric / Number of words this lyric) * ln(Total number of lyrics / Number of lyrics with word x)

The measurement type for the K-means clustering was nominal and the numerical measurement was cosine similarity (Lee & Bradlow, 2011). With cosine similarity, RapidMiner tries to maximize the cosine between two word vectors (Tata & Patel, 2007). When the angle between the two is zero, the cosine becomes one. The closer the two word vectors are, the closer to one the cosines will be. Based on this, RapidMiner can group them together in the predetermined number of clusters. Finally, the number of maximum runs was set to 5.

5.4. SPSS Analysis

The last step for each method is a regression analysis in SPSS 20.0. A linear regression was performed with all the relevant themes or other variables. Before performing the analysis, the assumptions for a linear regression (linearity, independence, homoscedasticity and normality) must be checked. In case any of these assumptions are violated, the implications as well as the data need to be reviewed.

6. Data - Process and Summary

This chapter will discuss the details on how the various variables were collected and some statistics about these variables. First the dependent and independent variables will be discussed, followed by the control variables.

6.1. Dependent and independent variables

The data set which has been used has been acquired through a variety of means. First, a selection of songs was made. This was done by taking all the songs which were in the weekly US Billboard Top 40 Singles Chart between April 8, 2006 and March 5, 2011. This resulted in a maximum possible data set of 893 songs. This number seems reasonable due to the trend that songs often remain in the chart for multiple weeks. The data, which were acquired from online software, called Tableau Software,[3] consisted of the single title and artist. Tableau Software allowed for a selection of a time period of songs from the US Billboard Top 40 Singles Chart and for downloading them in an Excel format. The website address of this online software has been added as a footnote.

The next step includes acquiring the lyrics. For this, a crawler was custom-written to acquire as many lyrics as possible with minimum effort. The latter is important to ensure that companies can easily repeat the process. The crawler acquired the lyrics from one source, namely the website lyrics.wikia.com and it was able to acquire the lyrics of 743 songs.

The follow-up step for this concerned the Last.FM data. Another crawler was created to acquire the data from the Last.FM API. The API was able to provide the play count of all 743 songs of which lyrics were included. Furthermore, the API was able to identify another 33 songs based on the title and artist. For these 33 songs, the lyrics were added manually; 7 could not be found on lyrics.wikia.com and therefore were extracted from various other lyric websites. This therefore resulted in a total data set of 776 lyrics with play count data.

For 117 songs from Last.FM there was insufficient data. After an investigation into the matter, it seems that all songs had small differences in possible song or artist names. This could be due to cultural differences although items such as a variety of versions of the specific song existed.

[3] http://public.tableausoftware.com/views/BillboardChartData_Top40/MasterDash?:showVizHome=no

Collecting this data could cause inconsistencies which could be systematic, while it is not believed that the current data set has systematic errors due to the missing songs. An example of an error could be that the album and radio versions are both named differently while they are very similar. The play count in Last.FM is therefore split up instead of combined, which gives the wrong image. Therefore, it was regarded safer to disregard these 117 songs from the analysis.

6.2. Control variables

Concerning the control variables, Last.FM was able to provide various tags, provided by the users, which were attributed to certain songs. When attributing tags through the Last.FM API, 28 lyrics did not receive any tags. Of these 28 songs, the tags could be determined for 21 songs through the Last.FM website.

To determine possible genres, all tags which were attributed to the songs were analyzed manually. Here, a wide variety of genres were extracted, which can be found in table 1 of Appendix A. Sometimes a mixed combination of two genres was found in a tag, such as "pop rock". Whenever this was the case, both genres were deemed part of the song; thus, both were coded.

Besides the possibility of one tag having multiple genres, one song also has multiple tags. Therefore it is possible for a song to receive multiple genres because in different tags a genre could be found.

Although the tags might not indicate the genre the song was originally produced for, it does indicate how the listeners experience and judge the song. Using the API to determine the genre is therefore a good way of crowd sourcing. If the public has a certain preference for specific genres, it is believed the Last.FM API is much more capable at showing this than the "official" genres.

The tags which were used for acquiring genres showed another interesting phenomenon. As stated above in the literature, it seems that female singers get higher sales in specific themes (Lewis, 1993). Since coding this by hand is not possible on larger data sets within the time constraints, this has not been done. The Last.FM tags, however, showed that there might be another way. It seems that the public is sometimes feministic or prefer female vocalists. This causes female vocalists to be tagged specifically as a female vocalist. Thus, using this as a dummy variable suddenly became very reasonable. It might not be possible to find all female vocalists, but it would be interesting to run an analysis with this control variable to see whether it is significant. Furthermore, the duration and listeners of the songs were acquired without any issues.

An issue that arose is the publication date of songs. The researcher tried to find a consistent release date by using the Last.FM API; however, the Last.FM API could only provide around 300 release dates. When trying to resolve this, a possible was to use the album release dates. Here, however, similar issues were found with varying availability of release dates.

After further investigation, it was found that the publication date of songs can vary widely. It is therefore impossible to determine whether the release dates, if they can be acquired altogether, are consistent across the full data set. Reasons for this are the different release dates in different countries, as well as whether the first release was on an album or single. The latter is important since the US Billboard Top 40 Singles Chart only includes singles, while Last.FM will count any release before that.

Finally, the natural logarithm of date of publication was regressed against the play count of the albums for which this information was available. This resulted in a significant explanatory power of the model to consider this important enough. Two options were considered here: the first was to exclude the songs which did not have this data available. The second was to use the average on all songs where this data was not available. The latter caused a significant decrease in explanatory power in the pilot analysis. Therefore, it was decided to drop all the lyrics which did not have an album publication date available in the Last.FM API since it was believed not to cause systematic errors. When the date of publication of an album was available, it was confirmed that the album was published no later than May, 2011. This choice was made so that a consistent date of publication could be guaranteed, while also giving the model the most optimal control variables.

After excluding data due to these two reasons, it is believed that the data can still be considered a random sample. The amount of excluded relatively popular and less popular songs were almost the same.

When considering the data set, it was found that around 100 lyrics did not have any genres attached to them. These were therefore excluded from the analysis. The final data set contains 435 lyrics.

7. Analysis

The initial analysis will consist of three major parts. First the process in RapidMiner and determining the optimal number of clusters will be discussed. Furthermore the contents of the various clusters and the SPSS analysis will be discussed.

To confirm the results of the initial analysis two additional checks will be done in the robustness check. Finally some discussion will be had about the initial analysis with the robustness checks in mind.

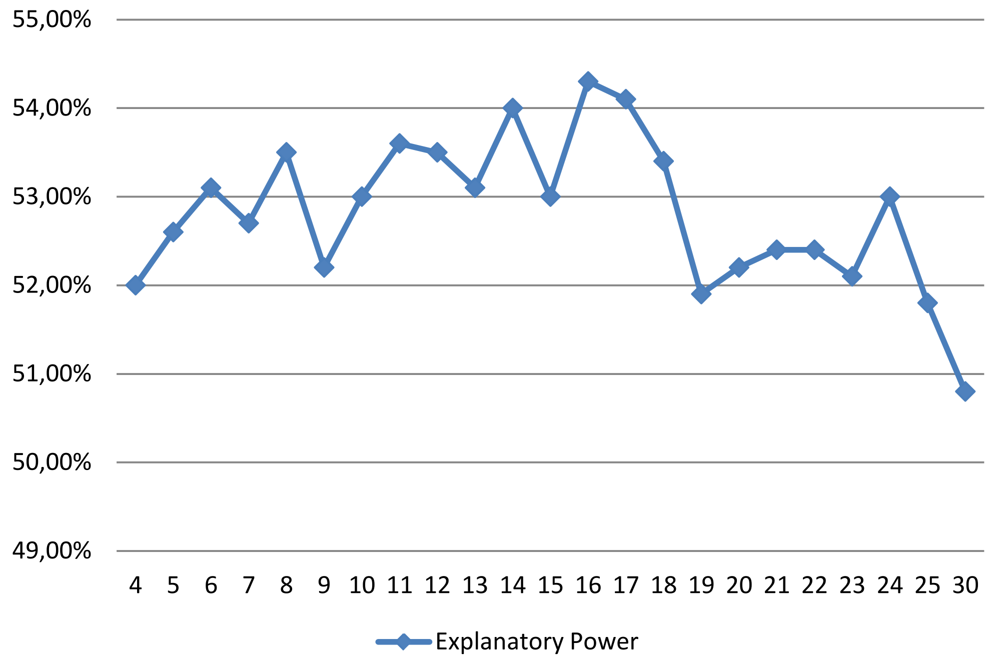
7.1. RapidMiner

As described before, the number of clusters is based on the adjusted R-square. To start the analysis, it is necessary to determine the number of clusters first. This has been done by running RapidMiner and the resulting clusters were used to create dummy variables in SPSS. In SPSS, all the control variables were used to determine the highest possible explanatory power. Therefore, all the genres, female vocalists, and the natural logarithms of artist listeners and duration were used.

When determining the number of clusters, the first 4 up to 25 clusters were considered. The goal was to determine the optimal point resulting in the highest explanatory power. The results can be found in graph 1 below. There seems to be a peak near cluster 16 with an adjusted R-square of 54.3%.

To confirm this, 30 clusters were also considered. The higher ones did not seem viable due to the time it took RapidMiner to complete 30 clusters. Furthermore, 30 clusters resulted in a comparatively low explanatory power. Therefore, not considering a higher amount of clusters is reasonable, besides the time constraints. With the results of the higher number of clusters in mind, it was confirmed that the optimal point was at 16 clusters.

Graph 1: Explanatory Power



7.2. Assumptions

The first assumption to be considered is the presence of outliers. Outliers in the data were determined by making a scatter plot. On the Y-axis the Regression Standardized Residual could be found, while the Regression Standardized Predicted Value could be found on the X-axis. In the scatter plot, 15 points were found which were below -3 on the Y-axis or X-axis. These were therefore removed from the results. When rerunning this regression, it was found that another 2 points became outliers; hence, these were also removed. This resulted in a final sample of 418 data points. This improved the explanatory power to 55.1%. The final scatter plot can be found as graph 2 in Appendix B.

When considering the normality assumption, graph 3 and 4 of Appendix B have been used. As can be seen, the normality assumption was not fully met. It was met, however, up to a certain extent. Since the linear regression is fairly robust when it comes to this assumption, it is still possible to use the linear regression (Lumley et al., 2002).

For the homoscedasticity assumption, the scatter plot used for outliers was considered. Since no real answer could be determined based on this, a Koenker-Basset test was performed. The preference of Koenker-Basset over Breusch-Pagan came from the issue that the Breusch-Pagan can give errors when the normality assumption is not fully met. The Koenker-Basset test gave a significance level of 0.7672 on the Chi-square, thereby supporting the assumption that the data is homoscedastic.

Furthermore, when the collinearity statistics are considered, it was concluded that none of the VIF statistics are above 2.5. Rogerson (2001) said that when the VIF is below 5, no multicollinearity issues exist. Therefore this assumption was met and no significant collinearity exists between the independent variables.

Finally, the linearity assumption was considered. As shown in graph 2 of Appendix B earlier, this does not seem to be a significant issue. The possibility of a non-linear model was minimized as soon as some of the variables were transformed into natural logarithms based on some pilot analysis.

7.3. Descriptive statistics

The first step in the analysis was to consider the descriptive statistics. This has been done to make sure that there are no discrepancies. While doing this, the first step was to look at the general statistics, as summarized in table 2 of Appendix A. While creating this table, it was initially discovered that the duration was showing few discrepancies. The minimum value in these cases was showing some unexpected results. While investigating these cases, it was found that three did not have the appropriate duration. It seems the low values were coming from thirty second samples. These were manually corrected, which gave the final result in table 2 of Appendix A.

In table 2 of Appendix A, it can be seen that the play count has a very low minimum but a relatively high maximum. Similar issues existed with the duration, e.g. the days since its release and the number of listeners of a specific artist. Therefore another confirmation that a natural logarithm will probably model a better fit.

In the same table, it can be seen that the mean of the play count seems to be relatively high. The mean was around 1.7 million. Due to the standard deviation of 2.1 million it can be argued that there are quite a few "extreme" performers. A similar concept can be applied to the listeners of an artist; here the standard deviation is much less compared to the mean though.

The days since its release were measured from the day of acquiring the Last.FM data. Therefore, the days since the release were measured based on the time since the album release and until May 1st, 2014. It was decided not to allow any songs to exist in the section which were released 1000 days ago prior to May 1st, 2014. Sometimes, it happens that the single was released before the album. This is why a few extra months of releases were allowed in 2011, before the period of which the US Billboard Top 40 Singles Chart data was collected. While doing this, this resulted in an average of 2023 days since release, which is slightly over 8 years ago, thus in 2006.

Therefore, it is probable that many songs which were in the US Billboard Top 40 Singles Chart were published before they were in the charts. This is possible because some songs might not be published as a single before the album has been released for quite some time. Similarly, some singles might become a hit again years later and show up in the charts for a second time. Therefore, it is reasonable to keep using the album release data as they were able to acquire listeners and play count on Last.FM.

The next step is to look into another control variable: genres. Since genres are presented to SPSS as dummy variables, it is a little bit more difficult to give useful information about them. In table 1 of Appendix A, all the genres can be found with the number of lyrics they were applied to. As visible in table 1, almost every genre comprises some lyrics. This is of course due to the fact that the genres were determined based on tags. The only exception is Jazz since it only had a frequency of 1 before the assumptions were checked. Since some other genres are showing very low frequencies, three or lower, these were disregarded in the analysis. Furthermore, the maximum is in the pop genre. 171 lyrics were determined to have an affiliation for this genre. This has to do with the fact that pop music is considered to be the main stream music which cannot be denied based on these frequencies. More surprising is the fact that it is not the only type of music with higher frequencies. Hip-Hop, Rock and R&B can also claim a large share. We must keep in mind that a song can be part of multiple genres, as described in the data chapter.

Furthermore, female vocalists were considered. The frequencies show that 108 songs were marked as having a female vocalist, which is 25.8% of the considered data set. This seems to be a decent estimate when considering the number of groups, duets and male vocalists being in the charts. It is therefore considered a control variable which can be used.

Finally, the frequencies of clusters are presented. These can be found in table 3 of Appendix A. Just as with the genres it is not possible to give detailed information. What can be seen is that every cluster still has lyrics in it after dropping part of the data set. The lyrics show a tendency to group in one cluster (c7). Therefore, this has been used as base category to avoid perfect multicollinearity. The minimum amount of lyrics in a cluster can cause issues with cluster 3, 5 and 6. These lyrics might not be able to show the actual effect due to the low frequencies present in the clusters. All other clusters contained enough lyrics.

7.4. Initial analysis

Since most of the assumptions were met, a linear regression was performed on the data. The linear regression results can be found in table 4. As described earlier, c7 was dropped from the analysis in SPSS to avoid perfect multicollinearity and to create a base category.

This table shows that few of the control variables performed very well. This includes the control variables artist listeners and the days since release. Contrary to earlier pilot results, duration does not seem to have an effect. When considering the genres it is visible that the results vary widely. The genres R&B, dance, electronic, indie, alternative and urban are all significant at a 95% confidence interval.

Most surprising was the fact that the most common genres in this sample, pop and hip-hop, do not seem to have a significant effect. This is probably due to the very diverse set of lyrics present in those genres; resulting in widely varying, unexplained, play counts. Therefore, it is difficult to control. Most other genres have a relatively low number of lyrics. Therefore, these possibly have widely varying play counts as well; thus, they cannot be controlled.

Finally, with regards to the control variables, it can be seen that female vocalists show a significant positive effect. Although the measurement here might not be perfect, it does shows that it is a significant contributor.

The next step of the analysis is the most interesting one. Since this research is interested in whether there is any predictive power in text analysis, it is important to know whether there are any effects present in the various clusters as determined by SPSS. When the variables c0 up to c15 are considered, it can be seen that quite a few show a significant effect. First, a quick overview will be given of the significant clusters and their meaning after which the effects will be discussed.

Table 4: Regression analysis

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-1.605	4.638		-0.346	0.730
ln_artist_listeners	0.829	0.054	0.538	15.329	0.000***
ln_days_since_release	0.564	0.214	0.100	2.636	0.009***
ln_duration	-0.104	0.343	-0.011	-0.302	0.763
genre_rap	-0.162	0.199	-0.033	-0.815	0.416
genre_rnb	-0.405	0.156	-0.098	-2.600	0.010***
genre_pop	0.100	0.137	0.030	0.731	0.465
genre_hiphop	-0.312	0.172	-0.080	-1.811	0.071*
genre_dance	0.554	0.177	0.123	3.127	0.002***
genre_party	0.290	0.599	0.017	0.485	0.628
genre_soul	-0.046	0.249	-0.007	-0.185	0.853
genre_electronic	1.021	0.296	0.131	3.447	0.001***
genre_rock	0.274	0.182	0.071	1.507	0.133
genre_indie	0.997	0.259	0.137	3.844	0.000***
genre_alternative	0.710	0.223	0.159	3.188	0.002***
genre_country	0.118	0.239	0.018	0.493	0.622
genre_ballad	0.204	0.359	0.019	0.568	0.571
genre_punk	0.471	0.392	0.044	1.203	0.230
female_voc	0.479	0.142	0.129	3.388	0.001***
c0	0.058	0.190	0.012	0.303	0.762
c1	0.996	0.469	0.073	2.122	0.035**
c2	0.458	0.206	0.087	2.227	0.027**
c3	-3.171	0.654	-0.165	-4.846	0.000***
c4	0.201	0.299	0.024	0.672	0.502
c5	0.444	0.648	0.023	0.685	0.494
c6	0.331	0.564	0.020	0.586	0.558
c8	0.988	0.383	0.088	2.578	0.010***
c9	0.311	0.446	0.025	0.697	0.486
c10	0.207	0.190	0.044	1.087	0.278
c11	0.236	0.330	0.025	0.717	0.474
c12	0.155	0.196	0.031	0.786	0.432
c13	-0.137	0.343	-0.014	-0.400	0.690
c14	0.567	0.230	0.093	2.467	0.014**
c15	0.337	0.301	0.040	1.122	0.262

Adjusted R-square: 55.1%

* 90% Significance, ** 95% Significance, *** 99% Significance

When considering the various clusters, the significance level is important. The significance level is based on the t-test which has been performed. This is based on the non-standardized coefficients. The hypothesis of this t-test is that the coefficient is equal to zero. Therefore, when the significance level is below 0.05, the coefficient can be considered significant. Resulting in the null hypothesis getting rejected, at a 95% confidence interval.

In general, the clusters with a significance level above 0.05 seem to have issues with their standard error which is relatively high in comparison with the estimated coefficient. Most of the coefficients are only one standard error away from zero, having a significant chance that the actual effect of these clusters is zero or not relevant. These clusters are therefore not significantly beneficial to the model.

When interpreting the results, it is important to keep in mind that c7 was left out, and is therefore the reference point of the whole data set. Furthermore, all coefficients were judged based on a 95% confidence interval.

7.5. Cluster analysis

The cluster analysis will focus on describing the various clusters. Since each cluster is unique in its contents it is important to know what each cluster consists of. Therefore first the reference cluster will be described, followed by the significant and insignificant clusters.

Reference cluster

Cluster 7 is the reference cluster of the analysis. The main driver of this cluster is the word "love". Other words that are common are "find", "baby", "heart" and "give". This cluster is focused on love and the words are easily combined with each other in this theme. Many lyrics were related to love and this cluster therefore contains 96 lyrics.

Significant clusters

Based on the significance levels of the regression analysis in table 4, some meaning can be given to the significant clusters. Since only five clusters were found to have significant values, these will be discussed first. This will be done in two ways: first, the cluster analysis of RapidMiner will be considered and second, the clusters will be evaluated based on the manual coding of the lyrics in that specific cluster. The manual coding was done by examining common believes of the song’s meanings. Specific websites were set up so people could share their thoughts on the meaning of a particular song. Many of these websites were used together with the authors' coding of the song to roughly determine the main themes. Finally, the significant clusters which will be discussed are c1, c2, c3, c8 and c14.

Cluster 1 contains lyrics which are clustered mostly on the words: "forever", "null", "ella", "umbrella", "telephone", "star", "raining", "young" and "calling". The songs which fit in this cluster can be found in table 5 of Appendix C. In searching for the meaning of the songs based on reviews, it seems that the consensus amongst these songs is that they are all love or sex related. Some of the words could fit this profile include "forever", "telephone", "young" and "calling". This theme focuses on the happier side of life, including love.

Cluster 2 is mostly clustered on the following words: "tonight", "yeah", "light", "free", "night", "hand" and "drive". The songs which are part of this cluster can be found in table 6 of Appendix C. One noticeable thing is that the majority of the songs here are also focused on love, possibly in combination with something else. It therefore seems that both cluster 1 and 2 contain words which are often connected to love. This cluster focused slightly more on nightlife in combination with love.

The third cluster is showing a connection with the words: "mum", "raise", "glass" and "kind". Together with the list of songs which were part of this cluster, see table 7 in Appendix C, it seems that the cluster is made up of songs which were put together by chance even though the significance was as low as possible. The random chance was possible because the number of lyrics in this cluster is only 3. Therefore, giving meaning to this cluster based on the given words was much harder.

The next cluster is c8. In cluster 8 the words "blame", "sorry", "sad", "late", "girl", "apology" and "bed" are the most common ones. In table 8 of Appendix C, the relevant songs and comments can be found. Although this cluster showed quite some relations towards love songs, it seems that the songs were of a different kind. Although the most common words are connected to love, the lyrics are of a sadder kind. The songs can therefore be considered more of the apologetic kind.

The final cluster is c14. Cluster 14 is a bit different compared to the clusters described before. This is not only because it is relatively big in size but also because the number of common words is relatively low. Words that are the most common include: "girl", "yeah", "baby", "kiss", "meet", "cold", "want", "tell" and "said". As can be seen, these words are more of the general type compared to the clusters described before. One interesting thing in this cluster is that it contains the song "Umbrella" by Rihanna, while cluster 1 shows connections with the word "umbrella". In table 9 of Appendix C, the full list of songs can be found. As can be seen in the same table is that the word "love" appeared often. The clusters also contains many songs about reflection.

When looking at the various clusters which have been discussed, it can be seen that love in its various forms is often present. There are two possible explanations for this. The first is that love related songs are more common in general, which is confirmed by the reference cluster. The second option is that these clusters specifically captured love niches. A robustness check has been done to find out whether love in general is more common, or whether it relies on specific words. These two insignificant clusters were manually coded. The first was relatively small in size while the other was larger. One thing which is important to realize is the fact that RapidMiner stemmed from all words. Therefore, the words presented above can have variations in the lyrics. For example, love, loved and loving are all considered to be part of the same stem.

Insignificant clusters

Besides the clusters which were described before, there were another 11 clusters which were not significant. Manual coding was being avoided in this thesis, except for gaining more insights in the clusters. However, this research found it important to give the relevant words per cluster and to do the robustness check. Therefore, first the relevant words per cluster will be summed up as before, after which for two random clusters one small and one large one, a robustness check for manual coding will be done.

The first cluster to be discussed is c0. In RapidMiner, it was shown that this cluster had no connections with any words Therefore, it can be concluded that this was the last cluster that has been formed, which contains the last remaining lyrics. An insignificant coefficient was therefore expected.

Cluster 4 consists of words such as "shake", "body", "move", "gimme", "higher" and "magic". Based purely on these words, there probably are probably some sexual or seduction related themes present in this cluster.

Cluster 5 is kind of special. Apparently, there are songs that have a very strong connection with the word "na", which is the main driver of this cluster. This cluster is, however, only driven by 3 lyrics. When manually checking the lyrics, it was found that "na" is indeed a very common word. It is often used to fill the gaps between parts of the lyrics or as an introduction.

The fourth cluster to be discussed is cluster 6. In this cluster, the words "burn", "rehab" and "disease" appear relatively often. This cluster can therefore be described as a slightly sad cluster, which is s focused on more serious subjects. This cluster is, however, only driven by 4 lyrics.

Cluster 9 showed some connections with "rock", "boom", "party" and "london". The lyrics in this cluster seem related to party or dance music. This cluster consists of 7 lyrics.

Cluster 10 was a mixed cluster without a well-focused set of words compared to some other clusters. This could be due to the size since it has 59 lyrics. The words which appear more often are "day", "sweet", "see", "I'm" and "love". This cluster shows similarities with cluster 7. It is not significantly different from cluster 7; thus, it cannot be considered a separate niche.

The following cluster is number 11. The main driver of this cluster was the word "oh" followed by "halo", "break", "go", "bad" and "smack". Similarly for the word "na" in cluster 5, this cluster showed similar behaviour with the word "oh".

The next cluster is cluster 12. Here, common word occurrences are "ya", "money", "shawty", "girl", "they" and "nigga". The words which were used in this cluster seemed to be more focused on a niche in the music genre. It is focused on money and girls.

Cluster 13 has correlations with the words "la", "loca", "gracenote" and "que". This cluster seems to focus on lyrics that are not in English. It is actually interesting to see that the language difference in general does not show a significant difference in popularity in the charts.

The final cluster to be discussed is cluster 15. In this cluster the most common words are "dream", "good", "something", "believe", "teenager", "anymore" and "late". This cluster shows correlations with more positive themes. It is therefore interesting to see that this cluster is insignificant even though the number of lyrics, 18, that is present in the cluster is reasonable. The words in general seem to relate to positive teenage related music.

7.6. Regression analysis

In this section the regression analysis is considered into more detail. The regression analysis gave five clusters which showed a significant coefficient at a 95% confidence interval. These will therefore be discussed in order of most significant to least significant.

The most significant cluster is c3. C3 had a negative sign of 3.171 on the natural logarithm of play count. The significance level is at 0.000. When a positive sign is present, it means that the play count on Last.FM is on average positively affected when a song is present in that specific cluster. The opposite holds for negative signs. When the complete regression is run, it is possible to give a prediction of a song based on the result from this linear regression.

When considering the number of lyrics in this cluster, 3, random chance cannot fully be excluded. This is less likely due to the fact that the lyrics were clustered by RapidMiner and manually coded independently of the listeners but random chance cannot be fully avoided with low frequencies. This was also one of the clusters in which it was hard to attribute a specific theme based on the words collected in the clustering.

The next cluster is c8. C8 has a positive sign of 0.998 on the natural logarithm of play count. The significance level is at 0.010. This cluster contains 9 lyrics were present, which is 2.2% of the total data set.

The third significant cluster is c14. C14 is showing a positive sign of 0.567 on the natural logarithm of play count. The significance level is at 0.014. In this cluster, 32 lyrics are involved, which is 7.7% of the data set.

The following cluster to be considered is c2. C2 shows a positive sign of 0.458 on the natural logarithm of play count. The significance level is at 0.027. In this cluster, 44 lyrics were present, which is 10.5% of the total data set.

The final cluster to be considered is c1. C1 shows a positive sign of 0.996. The significance level is at 0.035. Similar to c3, this cluster does not have a huge number of lyrics, although it is twice the size of c3.

7.7. Robustness

The following step of the analysis is the robustness check. One of the things that were noticed in the significant clusters was the fact that many songs are somehow related to love. Whether this is more a general trend in chart music should be checked. To achieve this, this research is going to manually code two more insignificant clusters for comparison. When deciding on which of the clusters these should be compared, a random number generator was used. The only insignificant cluster which will not be included in this number generator is cluster 7. This is the reference cluster which already shows correlations with love. The random number generator determined that cluster 12 and 15 will be considered for the robustness check.

When checking the various lyrics in cluster 12, which can be found in table 10 of Appendix D, it is found that there are less love related lyrics present than in other clusters. This is probably due to the type of words which were found. The cluster seems to focus on a broader set of themes compared to the clusters with significant results.

Cluster 15 does not show a similar pattern to the results found in the significant clusters. This can be easily linked to the words that seem to focus more on being young. The list of songs in cluster 15 can be found in table 11 of Appendix D.

Therefore it can be argued that words are able to capture specific themes within the music relatively well, since it was able to distinguish less and more love related clusters.

As an additional robustness check, it had to be made sure that no significant correlations exist between the genres and clusters. To achieve this a principal component analysis ("PCA") was conducted.

The most important thing to note here is that genres which contained three or less lyrics were not included in the analysis, because three or less lyrics were not considered representative of a specific genre. The same condition has also been used in the regression analysis. The data used for the principal component analysis is once again from two different data sources. The clusters were created by RapidMiner based on the lyrics, while the genres were acquired with the Last.FM tags.

In the next paragraph, a strong correlation is described when the achieved matrix value is between (-)0.5 and (-)1, while a weak correlation describes a value between (-)0.3 and (-)0.5.

In the varimax rotated component matrix found in table 12 of Appendix D, it can be seen that the factor which has a strong correlation with c13 shows strong correlations with hip-hop and rap and a weak correlation with pop. Similarly, the factor which has a strong correlation with c14 has a weak correlation with c13, soul and R&B. While the factor which has a strong correlation with c2 shows a weak correlation with pop and the factor which has a strong correlation with c15 shows a strong correlation with country and a weak correlation with c13. The factor which has a strong correlation with c5 shows a weak correlation with c13 and the factor which has a strong correlation with c6 shows a weak correlation with hip-hop.

Based on the rather limited number of strong correlations in the factors, it is unlikely that specific word combinations can only be attributed to a few specific genres. This is especially true when only the significant clusters are considered, since these only show weak correlations. Rerunning the regression without the genres did not cause other clusters to suddenly become significant either. It therefore seems that the regression analysis was able to capture the effects properly.

7.8. Discussion

When keeping the robustness check in mind it can be said that there are certainly some interesting results. The most interesting result is the fact that a cluster which is focused on the word love and a few other words, is probably too broad. This can be seen in reference cluster, cluster 7, which was the general love cluster. This cluster showed no significant difference with most of the clusters in the regression analysis. Therefore, cluster purely about love is not enough to show the differences.

Keeping this in mind, it can be argued that there might be niches within this theme that fare better. This is where the significant clusters come in. This thesis was able to identify five clusters which did show significant results. However, one of the issues was that the clusters were showing quite high correlations with love in the manual coding based on websites with people's opinions about lyric interpretation. The robustness check found that cluster 15 shows a completely different word set which does not show high values of love in its lyrics. It is therefore fair to assume that the words which were found in these significant clusters, which cluster 15 was not part of, can be considered niches of the love theme.

The only exception to this is cluster 3. Cluster 3 only contained 3 lyrics. This therefore significantly skewed the results due to the possibility of random chance. Therefore, it cannot be argued whether this cluster is actually useful. This is chance is minimized due to the fact that the lyric and play count data were acquired completely separate from each other, but still present.

When looking at the significant clusters, a raw evaluation about how serious the songs in general are can also be seen. When comparing cluster 1, which has words such as "telephone", "umbrella", "star", "young" and "calling" in it, with cluster 8, which has words such as "blame", "sorry", "sad" and "apology" in it, it can easily be argued that cluster 8 is probably focused on more serious and apologetic songs.

7.9. Conclusion

In conclusion, it can be said that there are 4 clusters that benefited the regression analysis. These are clusters 1, 2, 8 and 14. These clusters have a high enough frequency to minimize the random chance of having only "good" or "bad" lyrics while at the same time they show a significant benefit in the regression analysis. All these clusters also show a positive sign towards the Last.FM play count; therefore, they are beneficial to the popularity of a song. All these clusters show a correlation towards love and, therefore, can all be considered niches in this theme.

8. Discussion and Limitations

In this chapter first a discussion will be had about the analysis and the literature is available. The second part will focus on the limitations of this research.

8.1. Discussion

This research has used a similar approach as Lee and Bradlow (2011). They showed that, based on clustering, it is possible to get information from product reviews. By reproducing some of the concepts they used in this research, it has proved to be viable method. Furthermore, it reinforces the importance of exploring this kind of methods in various text analysis situations. While setting up this research, various other techniques were considered but all seemed less viable compared to clustering.

Two main other options were considered. The first was the associate rules framework. This would allow certain word combinations to be formed by RapidMiner. This could then be used to code the individual lyrics for each of these word combinations. Another option was taking the most used words and use these as a means of dimension reduction such as principal component analysis in order to group them together. Although both techniques did not seem viable, they did introduce the best way to prepare the data since this was not mentioned in the paper by Lee and Bradlow (2011). One of the books which helped with this was written by Hofmann and Klinkenberg (2013).

When preparing the results, a few things can come into play. One thing that is noticeable while manually coding the lyrics is the depth of their meanings. Therefore, it is difficult to comprehend the actual meaning of the lyrics. This is supported in the literature by Greenfield et al. (1987), Hansen and Hansen (1991) and Hall (1998). Even though a lot of the lyrics revolve around love, it could be that people do not specifically interpret the songs in that way.

The manual coding was a way to confirm the results of another research which was performed by Hobbs and Gallup (2011). Although this research did find specific clusters without many reproductive messages in them, these were insignificant. The opposite showed in some significant clusters: the possibility of being more successful as suggested by Hobbs and Gallup (2011) with reproductive related messages. The one exception here was that there was a cluster which had love as the main driver. This specific cluster was not found to be significant. It is therefore arguable that within this specific theme, there are certain niches which perform better or worse. Since Hobbs and Gallup (2011) used the chart itself as a reference of success, this research was able to confirm part of their results by using a different measurement.

Although this research did not specifically examine the control variable g. female vocalist, it was found to be significant. It is therefore interesting to compare it with the research by Lewis (1993). He said that women perform better when they sing about certain themes. Although this effect is averaged out in this research, it could support his explanation

When it comes down to the managerial implications of this work, it can be seen that love can be a very beneficial theme in a song. If the song fits one of the significant clusters described before, they are more likely to become popular. Even when the love song does not fit into the successful clusters, there is no indication that they will perform, on average, worse than any other song.

The only precaution which must be taken is to determine whether this is still the case nowadays. The method has been introduced and is relatively easy to repeat for any other researcher or company if they want to determine the current popularity of recent songs. The data set is not very old but it might be that people became a little bit annoyed with specific niches in love songs and are therefore shifting their preferences.

The scientific implications of these songs are much more interesting, mostly because it opens up a lot of options for future research. This research was able to pioneer the method in the lyrics area while at the same time confirming some of the concepts and ideas discussed in the limited available literature. Furthermore, this research gave the opportunity to show that automated content analysis can also be applied to non-user generated content. The lyrics varied widely, while the sample set was not as big as most user-generated content samples. It showed the viability of content analysis in another area of research and the possibility of using it in other fields as well. More ideas for future research will be discussed in the last chapter.

8.2. Limitations

The main limitation of this research is the fact that words have a context. Similar words can be used in very different contexts and therefore result in different types of songs being compiled in one cluster. This can already be seen in the created clusters . During the manual coding, there were lyrics about love, with all kinds of different themes within love. In the same cluster, there were lyrics about world change and being famous. This simple example shows how difficult it can be to cluster similar songs together. The fact that the manual coding of the lyrics was found to be mostly directed at one theme, is interesting as it increases the reliability of the results.

A way to increase the reliability and resolve the context issues is by looking into the possibility of using a bag-of-sentences. Instead of relying on word counts in a document, whole sentences can be considered. For this, it is necessary to increase the sample size significantly and to define the filters for words. Both are able to break or make this method. If it works, however, it would be very interesting to see the results.

Content analysis of which clustering is a part, is a means of acquiring information on what is included in the text. This is therefore one of the limitations of this research since it does not consider why the text is present in the document. If the question would be asked why all these lyrics have these subsets of words in them or why people like this kind of songs, this research cannot give an answer.

Furthermore, due to the setup of this research, it is only possible to acquire certain lyrics of songs. Copyright holders, unpublished lyrics of songs and other disturbances might limit the possibilities of acquiring a representative sample. Therefore, limitations in the subset of songs which can be tested and acquired and will always be present.

Another limitation concerns the data collection. The data collection was tried to be kept as consistent as possible. However, it has its limitations. One, it would have been possible for HTML-tags to end up in the sample. To make sure this did not happen, some additional precautions were taken in the crawler and a small random sample was taken to manually check the data. Finally, the RapidMiner results were also checked to minimize this possible issue.

Similar to the HTML-tags, recognizing additional words which had to be filtered did not go exactly as planned. To do this, the list of most frequently found words was used; however, when the clusters got explained in the analysis two more or less fillers came through the whole process unnoticed. Therefore, the clusters evolving around "na" and "oh" were part of the limitations of this work. These should not have been included as words in the cluster analysis. Unfortunately due to time constraints, this could not be addressed properly by redoing the clustering.

Furthermore, over the course of compiling the data set, about half of the data set had to be excluded at some point. During various moments in the research, it had to be checked whether it was still random or whether a consistent pattern could be detected. One of the methods used for this was to see how many songs per cluster, due to various reasons, were excluded compared to the total set. The only cluster which this was outside of the 45-55% of the songs dropped was cluster 3. Since this cluster was already considered small in terms of frequencies, no conclusions were drawn from this cluster.

Finally, Last.FM provided a means of acquiring consistent data on a variety of variables. Therefore, the trust in this data source must be put here as a limitation as well. This thesis depends not only on the fact that people scrobble their listening to Last.FM, but also on the concept that users share their opinions through tagging. Although this is not a perfect sample, is it hoped that by using crowd sourcing these effects were minimal.

9. Conclusion

This research investigated the possibilities of using automated text analysis as a measurement of popularity. While performing this research, the US Billboard Top 40 Singles Charts between April 8, 2006 and March 5, 2011 have been used to determine which lyrics to use. For the measurement, the play count of Last.FM was used for a linear regression. To determine the clusters, k-means clustering in RapidMiner was used. The final result of RapidMiner resulted in 16 clusters.

The linear regression showed that many of the effects were very well captured by the control variables. Mostly, the popularity of the artist and days since release were both very good additions while the duration did not add value. Furthermore, some of the genres also benefited the regression analysis. Finally, inspired by research of Lewis (1993), female vocalists were also confirmed to have a significant effect.

When it comes down to the themes, it can be seen that there are significant clusters. Almost all of the significant clusters contained a number of lyrics that gave a decent estimate of linear regression. The only exception here was the third cluster. Clusters 1, 2, 8 and 14 all showed significant results with at least 6 lyrics. When summing up the various clusters, it was clearly visible that each cluster had its own unique words. This showed the power of this method. Finally, 54 out of 418 lyrics did not have associations with any cluster.

The significant clusters with a high enough frequency all showed a positive sign towards the play count. When manually coding these clusters to figure out a general trend towards specific themes, it was found that most of the songs were love, sex or reproductive related. One of the insignificant clusters, however, showed a general trend towards love songs. Hence, this showed the possibility of niches within this theme which might explain the various rates of success better.

With this in mind, the research question, *which themes are more likely to become popular songs based on its lyrics*, can be answered. The answer is that specific niches within the love themes are able to create more popular songs in the long run between 2006 and 2011. The latter is important to note because Last.FM is a measurement of popularity in the long run. The US Billboard Top 40 Singles Chart is also a popularity measurement; thus, all these songs already made it to that specific point.

The cluster which causes the highest effect is number 1, which has the words "forever", "null", "ella", "umbrella", "telephone", "star", "raining", "young" and "calling" associated to it. This can roughly be summed up as lyrics about love and happiness. The clusters which are most successful after this, in order, are number 8, apologetic lyrics, 14, reflection related lyrics, and 2, love and nightlife related lyrics.

10. Future Research

This research was able to identify some effects which are present by using k-means clustering. This method has certain benefits but also some downsides. As can be seen, the words to describe each cluster can sometimes be out of context. Furthermore, there were two clusters that were mostly based on a single word, which is not optimal. It would therefore be interesting to see whether there are other methods available which could reproduce these results. Other methods might also be able to capture concepts such as context of sentences better. While preparing for this research, the association rules framework was considered (Agrawal et al., 1993; Han et al., 2000; Tan et al., 2005; Hofmann & Klinkenberg, 2013) but it was unable to capture the lyrics properly. One of the solutions which might be able to resolve this is using a bigger data set.

Furthermore, it would be interesting to see how the lyric analysis would fare with a different dependent variable. Although, as described before, this research tried to capture the lyrics as realistic as possible, it would be interesting to see whether under different platforms the predictive power is significantly different. Examples of other dependent variables would be the advertisement budget, Itunes sales, CD or DVD sales and so on. Even though they all have issues with availability or reliability, the limitations must be recognized.

Besides these suggestions, improving the control variables would help as well. This research chose to keep the control variables as consistent and acquirable as possible but due to time constraints, the choice was to drop a significant portion of the sample. With manual labour, it would be possible to improve the completeness of the whole data set. This would therefore improve the research study. Similarly, some of the issues raised in this research need to be resolved, such as whether to use the first album or single release date, which country to use for the release date or whether to use the first release ever.

Finally, this research used the US Billboard Top 40 Singles Chart to select the lyrics, which focuses on US spins on the radio. This was done to make sure the lyric language would be as consistent as possible; thus, it was mostly in English. It would however be interesting to see how language influences popularity in various countries or whether using native languages can be predicted just as well as this research shows with the English language.

11. References

Agrawal, R., Imielinski, T. & Swami, A. (1993). "Mining association rules between sets of items in large databases." *ACM SIGMOD Record*, 22 (2), 207-216.

Atkinson, C. "Global music sales decline 4% by 2013. *New York Post*. March 18, 2014." *NYpost.com,* http://nypost.com/2014/03/18/global-music-sales-decline-by-4-in-2013/ Accessed June 8, 2014.

Carey, J.T. (1969). "Changing Courtship Patterns in Popular Song." *American Journal of Sociology*, 74 (6), 720-731.

Constance, B. & Silverblatt, R. (2014). "Sustainability Content Analysis of Management Texts." *Journal of American Academy of Business*, 19 (2), 59-65.

Elberse, A., Eliashberg, J. & Villenueava, J. (2005). "Polyphonic HMI: Mixing Music and Math." *Havard Business School Casy Study # HBS-9-506-009.*

Feldman, R. & Sanger, J. (2006). *The Text Mining Handbook*. New York: Cambridge University Press.

Freudiger, P. & Almquist, E.M. (1978). "Male and female roles in the lyrics of 3 genres of contemporary music." *Sex Roles*, 4, 51-65.

Greenfield, P.M, Bruzzone, L., Koyamatus, K., Satuloff, W., Nixon, K., Boride, M. & Kingsdale, D. (1987). "What is the rock music doing to the minds of our youth? A first experimental look at the effects of rock music." *American Journal of Psychology,* 24, 533-544.

Hall, P.D. (1998). "The relationship between types of rap music and memory in African American children." *Journal of Black Studies,* 28 (6), 802-814.

Han, J., Pei, J. & Yiwen, Y. (2000). "Mining frequent patterns without candidate generation." *ACM SIGMOD Record*, 29 (2), 1-12.

Hansen, C.H. & Hansen, R.D. (1991). "Schematic information processing of heavy metal lyrics." *Communication Research,* 18, 357-369.

Hobbs, D.R. & Gallup, G.G. (2011). "Songs as a Medium for Embedded Reproductive Messages." *Evolutionairy Psychology*, 9 (3), 390-416.

Hofmann, M. & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman and Hall/CRC.

Johnstone, J. & Katz, E. (1957). "Youth and popular music: A study in the sociology of taste." *American Journal of Sociology*, 62 (6), 563-568.

Knobloch-Westerwick, S., Musto, P. & Shaw, K. (2008). "Rebellion in the Topic Music Charts: Defiant Messages in Rap/Hip-Hop and Rock Music in 1993 and 2003." *Journal of Media Psychology: Theories, Methods and Applications*, 20 (1), 15-23.

Lee, T.Y. & Bradlow, E.T. (2011). "Automated Marketing Research Using Online Customer Reviews." *Journal of Marketing Research*, 48 (5), 881-894.

Lewis, G.H. (1993). *All that glitters: country music*. Bowling Green: Bowling Green State University Popular Press.

Ludwig, S., Ruyter, K. de, Friedman, M., Brüggen, E.C., Wetzels, M. & Pfann, G. (2013). "More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates." *Journal of Marketing*, 77 (1), 87-103.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). "The Importance of the Normality Assumption in Large Public Health Data Sets." *Annual Review of Public Health*, 23, 151-169.

Ramos, J. (2003). "Using TF-IDF to Determine Word Relevance in Document Queries reference." *Proceedings of the First Instructional Conference on Machine Learning*.

Rogerson, P. A. (2001). *Statistical Methods for Geography.* London: SAGE Publications Ltd.

Schwarz, N. & Clore, G.L. (2007). "Feelings and Phenomenal Experiences." *Social Psychology: Handbook of Basic Principles,* 2nd edition, New York: Guilford Press, 385-407.

Tan, P., Steinbach, M. & Kumar, V. (2005). *Introduction To Data Mining*. Boston: Addison-Wesley Longman Publishing Co.

Tata, S. & Patel, J.M. (2007). "Estimating the selectivity of td-idf based on cosine similarity predicates." *SIGMOD Record,* 36 (2), 7-12.

Wathieu, L. & Friedman, A. (2005). "Intelliseek." *Harvard Business School Case Study #505061-PDF-ENG.*

Appendix A: Descriptive statistics

Table 1: Genre descriptive statistics

Genre	Number of lyrics	Percentage of total
Rap	50	12,0%
R&B	79	18,9%
Pop	160	38,3%
Hip-Hop	92	22,0%
Dance	64	15,3%
Party	4	1,0%
Soul	26	6,2%
Electronic	19	4,5%
Reggae	2	0,5%
Rock	95	22,7%
Indie	22	5,3%
Alternative	65	15,6%
Country	29	6,9%
Ballad	10	2,4%
Punk	10	2,4%
Metal	2	0,5%
Latin	2	0,5%
Folk	2	0,5%
Jazz	0	0,0%
Techno	1	0,2%
Trance	1	0,2%
Urban	2	0,5%
Blues	1	0,2%

Table 2: General descriptive statistics

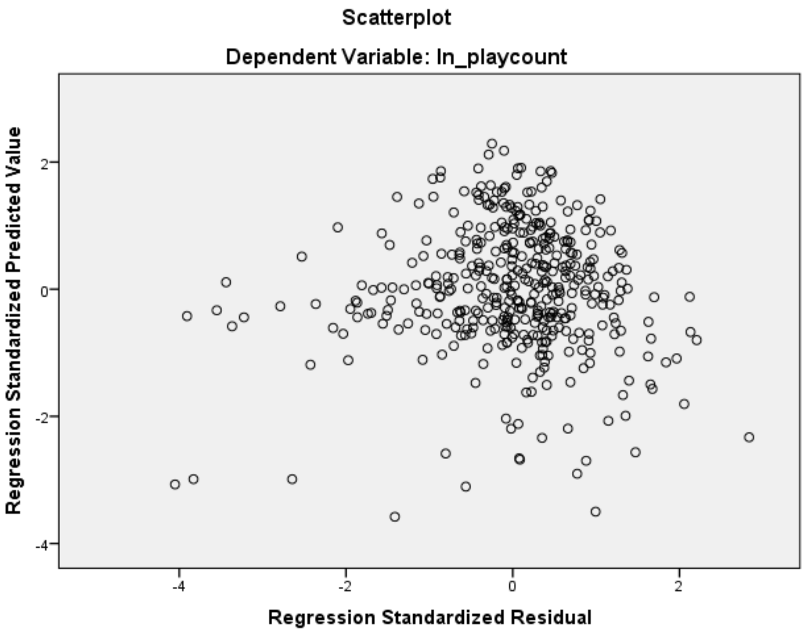
	N	Minimum	Maximum	Mean	Std. Deviation
Playcount	418	193	13044897	1702250	2135435
duration (miliseconds)	418	125000	448000	235806,2	38149,94
days_since_release	418	1025	3805	2049,2	585,6
artist_listeners	418	2785	4851922	1554655	1107171
Valid N (listwise)	418				

Table 3: Cluster descriptive statistics

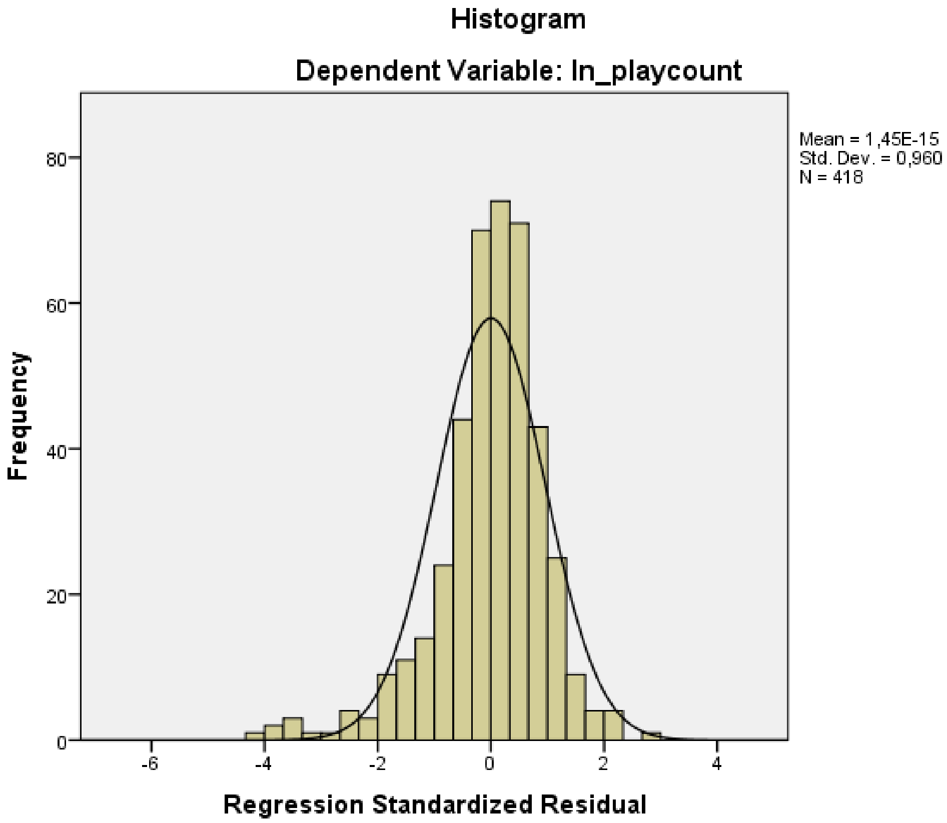
Cluster	**Number of lyrics**	**Percentage of total**
0	54	12,9%
1	6	1,4%
2	44	10,5%
3	3	0,7%
4	17	3,8%
5	3	0,7%
6	4	1,0%
7	92	22,0%
8	9	2,2%
9	7	1,7%
10	57	13,6%
11	13	3,1%
12	50	12,0%
13	12	2,9%
14	32	7,7%
15	16	3,8%
Total	418	100%

Appendix B: Assumptions

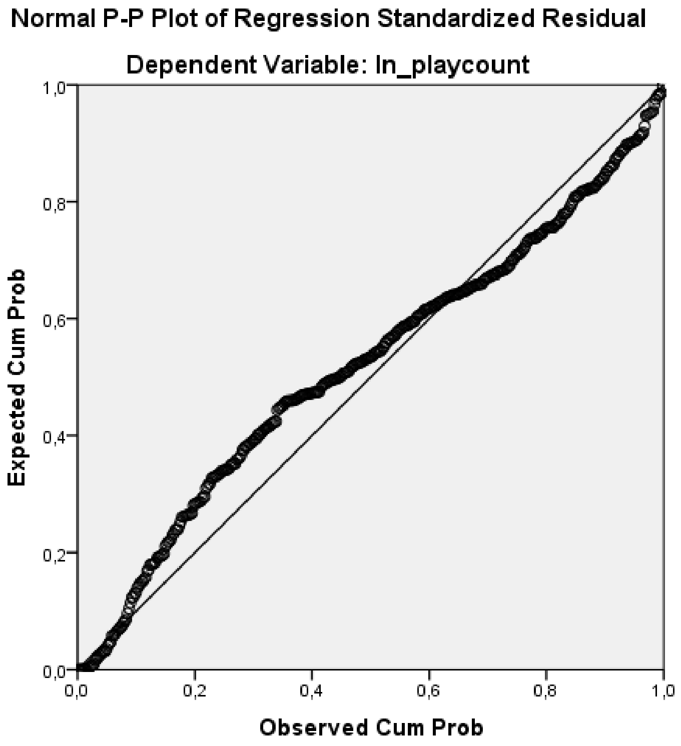
Graph 2: Scatter plot - Outliers



Graph 3: Histogram - Normality



Graph 4: Normal P-P Plot



Appendix C: Analysis

Table 5: C1

Artist	Title	Manul coding
Justin Timberlake	SexyBack	Sex & S&M
Ludacris	Money Maker	Sex
Lady GaGa	Bad Romance	Love & Relationships
Keri Hilson	Knock You Down	(Falling in) Love
La Roux	Bulletproof	Love & Break-ups
Jordin Sparks	Battlefield	Love & Fighting

Table 6: C2

Artist	Title	Manual coding
Rihanna	Take A Bow	Love & Cheating
Alicia Keys	No One	Love (declaration)
Fergie	London Bridge	Being famous
Akon	I Wanna Love You	Sex
The Black Eyed Peas	Imma Be	Breaking free from what you are
Kanye West	Stronger	Love & Seduction
Timbaland	Apologize	Love & Breakups
Train	Hey, Soul Sister	(Finding) Love
The Pussycat Dolls	Buttons	Love & Seduction
Colbie Caillat	Bubbly	Love
Yael Naim	New Soul	Enjoying life
Fabolous	Make Me Better	Love
Rihanna	Hard	Love & Breakups
Pink	Who Knew	Love & Breakups
Demi Lovato	This Is Me	Breaking free from what you are
Taylor Swift	Jump Then Fall	Love
The Pussycat Dolls	I Hate This Part	Love & Breakups
Carrie Underwood	Cowboy Casanova	Love & Seduction
John Mayer	Say	Love
Ashley Parker Angel	Let U Go	Love & Breakups
Neon Trees	Animal	Love & Relationships
The Pussycat Dolls	Beep	Love
John Mayer	Waiting On The World To Change	World change
Drake	Over	Being famous
Sean Kingston	Me Love	Love & Breakups
Playaz Circle	Duffle Bag Boy	
Nelly Furtado	Maneater	Love & Relationships
The Veronicas	Untouched	Love & Relationships
Rick Ross	The Boss	
The-Dream	Shawty Is A 10	

Akon	Beautiful	Love & Relationships
Maroon 5	Wake Up Call	Love & Relationships
Chris Brown	I Can Transform Ya	Love & Money
Ashlee Simpson	Invisible	Being famous
Justin Bieber	Never Let You Go	Love
Taylor Swift	The Other Side Of The Door	Love & Cheating
Kelly Clarkson	Because Of You	Life changes
Duffy	Mercy	Love & Cheating
Jeremih	Down On Me	
Keri Hilson	Pretty Girl Rock	
Hinder	Better Than Me	Love & Breakups
Augustana	Boston	Life changes
Glee Cast	Faithfully	Love
OneRepublic	All The Right Moves	World change

Table 7: C3

Artist	Title	Manual coding
Justin Timberlake	What Goes Around...Comes Around	Cheating
Lil Wayne	Gonorrhea	Sex
Jack Johnson	You And Your Heart	Life changes

Table 8: C8

Artist	Title	Manual coding
Katy Perry	I Kissed A Girl	Sexuality
Taylor Swift	You Belong With Me	Love & Friendship
Young Money	Bedrock	Sex
Drake	Find Your Love	Love & Friendship
Jordin Sparks	Tattoo	Love & Breakups
Young Jeezy	Put On	
Justin Bieber	Never Say Never	Life changes
Taylor Swift	Picture To Burn	Love & Breakups
Adam Lambert	If I Had You	Love

Table 9: C14

Artist	Title	Manual coding
Kings Of Leon	Radioactive	Reflection
Reba	Consider Me Gone	Love & Breakups
Iyaz	Solo	Love & Breakups
Kenny Chesney	Somewhere With You	Love & Breakups
Uncle Kracker	Smile	Love
Taio Cruz	Higher	Seduction
Glee Cast	Like A Prayer	Love & Religion
Red Hot Chili Peppers	Snow ((Hey Oh))	Reflection
Rihanna	Rehab	Addiction
Glee Cast	Total Eclipse Of The Heart	Love & Relationships
Robin Thicke	Lost Without U	Love & Relationships
Edward Maya & Vika Jigulina	Stereo Love	Love
Kelly Clarkson	Walk Away	Love & Relationships
Leona Lewis	Better In Time	Love & Breakups
Mariah Carey	Obsessed	
Akon	Right Now (Na Na Na)	Love & Breakups
Danity Kane	Show Stopper	
Katy Perry	Waking Up In Vegas	Enjoying life
My Chemical Romance	Welcome To The Black Parade	Love & Death
Plies	Shawty	Sex
Taylor Swift	Love Story	Love & Relationships
Dixie Chicks	Not Ready To Make Nice	Politics
The Fray	How To Save A Life	Friendship & Death
Timbaland	The Way I Are	Love & Relationships
Mariah Carey	Touch My Body	Seduction & Sex
Britney Spears	Hold It Against Me	Seduction & Sex
Lady GaGa	Born This Way	Be who you are
Taio Cruz	Dynamite	Enjoying life
Justin Timberlake	My Love	Seduction
Chris Brown	Kiss Kiss	Love & Breakups
Eminem	Love The Way You Lie	Love, Relationships & Breakups
Rihanna	Umbrella	Love

Appendix D: Robustness

Table 10: C12

Artist	Title	Manual coding
Usher	Love In This Club, Part II	Sex & Seduction
Birdman	Pop Bottles	Money
The Pussycat Dolls	Wait A Minute	Love
Lil Jon	Snap Yo Fingers	
T-Pain	Bartender	
Lil Wayne	Mrs, Officer	Seduction
Jay-Z	Young Forever	Growing up
50 Cent	Straight To The Bank	Money
Fat Joe	Make It Rain	Money & Strippers
The All-American Rejects	It Ends Tonight	Friendship
Rob Thomas	Her Diamonds	Love & Sickness
Mariah Carey	Bye Bye	Parents & divorce
Green Day	21 Guns	Military
Britney Spears	Piece Of Me	Being famous
Brooke Hogan	About Us	Being famous
Jessica Simpson	A Public Affair	Party & Enjoy life
Chris Brown	Gimme That	Seduction
Avril Lavigne	Keep Holding On	Never give up
OK Go	Here It Goes Again	Love & Sex
Fergie	Clumsy	Love
R. Kelly	Same Girl	Love
Glee Cast	Hello	Love
Keyshia Cole	Heaven Sent	Love & Breakups
The Black Eyed Peas	Boom Boom Pow	
Rihanna	Shut Up And Drive	Sex
The Pussycat Dolls	When I Grow Up	Being famous
David Cook	Light On	Love
Ludacris	How Low	Seduction
Colbie Caillat	Realize	Love & Friendship
Kanye West	Good Life	
Young Jeezy	I Luv It	Money
Saving Abel	Addicted	Love & Sex
Kelly Clarkson	My Life Would Suck Without You	Love
Stone Sour	Through Glass	Alchoholism
Jazmine Sullivan	Bust Your Windows	Love & Breakups
Finger Eleven	Paralyzer	Seduction
Jennifer Hudson	Spotlight	Love & Relationships
Jay-Z	Empire State Of Mind	New York

Taylor Swift	Fearless	Love
Justin Bieber	Somebody To Love	Love
Eminem	Beautiful	Addiction & Rehab
Shakira	She Wolf	Relationships & Seduction
Nelly Furtado	Promiscuous	Sex
Katharine McPhee	Over It	Love & Breakups
Rascal Flatts	What Hurts The Most	Love & Relationships
Secondhand Serenade	Fall For You	Love & Breakups
Jibbs	Chain Hang Low	Civil war
Lil Mama	Lip Gloss	Confidence
Kevin Rudolf	Let It Rock	Family relationships
Wiz Khalifa	Black And Yellow	Car

Table 11: C15

Artist	**Title**	**Manual coding**
Ashanti	The Way That I Love You	Love & Relationships
Lil Wayne	Got Money	Money
The All-American Rejects	Move Along	Friendship
Daughtry	September	Love & Breakups
Taylor Swift	Superstar	Love
Paul Wall	Girl	Love & Friendship
Ne-Yo	Sexy Love	Seduction
Glee Cast	Defying Gravity (Glee Cast Version)	Doing that you want
Far*East Movement	Like A G6	Money
Rihanna	Disturbia	Love & Breakups
Yolanda Be Cool	We No Speak Americano	
Jimmy Wayne	Do You Believe Me Now	Love & Breakups
Jessie James	Wanted	Love
Coldplay	Violet Hill	Politics
Shontelle	Impossible	Love
Kid Rock	All Summer Long	Love

Table 12: Principal component analysis

Rotated Component Matrix									
	Component								
	1	2	3	4	5	6	7	8	9
c0	,049	-,001	,006	-,914	,081	-,108	-,073	-,108	-,077
c1	-,012	-,014	-,015	,012	,008	-,039	-,043	-,019	,022
c2	-,052	,007	,020	,081	,044	-,066	-,078	,934	-,083
c3	-,023	,004	,019	,001	,012	-,005	,004	-,013	-,001
c4	-,004	,016	,032	,039	,041	-,035	-,008	-,038	,007
c5	-,015	,003	,056	,018	,053	-,015	,045	,000	,054
c6	-,025	-,004	,026	,008	-,007	,002	-,013	-,045	-,045
c7	,076	-,001	-,077	,467	,383	-,399	-,333	-,358	-,230
c8	,004	-,007	-,020	,064	,003	-,044	-,049	-,049	-,026
c9	-,017	,022	,026	,045	,011	-,057	-,015	-,028	,054
c10	,033	-,021	-,039	,059	-,938	-,074	-,081	-,064	-,062
c11	-,002	-,054	,006	,029	,064	-,024	-,013	-,027	-,030
c12	,035	,007	,000	,094	,082	,949	-,073	-,081	-,045
c13	-,152	,100	,474	,009	,096	-,073	,324	,010	,302
c14	-,046	-,023	-,195	-,019	,078	-,074	,781	-,041	-,089
c15	-,032	-,020	-,073	,045	,044	,021	-,074	-,052	,789
genre_rap	,165	-,047	,746	-,061	,037	-,018	-,091	-,013	-,032
genre_rnb	,260	-,238	,093	,208	,034	-,020	,403	-,028	-,169
genre_pop	,255	,001	-,343	,030	,134	-,053	,030	,351	,145
genre_hiphop	,108	-,167	,735	,056	-,016	,037	-,089	,024	-,123
genre_dance	,370	-,280	-,045	-,137	,215	,135	,043	,089	-,132
genre_party	,700	-,039	,102	-,086	-,030	-,022	,029	-,037	-,020
genre_soul	,407	-,055	,067	,147	-,093	,103	,353	-,091	,057
genre_electronic	,642	-,064	-,024	-,118	,207	,035	-,098	-,040	,003
genre_rock	-,074	,813	-,178	-,011	,066	-,047	-,068	,046	-,028
genre_indie	,470	,411	,038	-,020	-,042	-,026	,052	-,023	,022
genre_alternative	,265	,842	-,034	,007	-,030	,066	-,029	-,030	-,063
genre_country	,382	-,098	-,084	,008	-,011	-,112	-,052	,002	,504
genre_ballad	,662	,099	,081	,081	-,106	-,094	,006	,055	,021
genre_punk	,607	,302	,051	,079	-,017	,154	,013	,023	,058

Table 12: Principal component analysis (continued)

Rotated Component Matrix								
	Component							
	10	11	12	13	14	15	16	17
c0	-,059	-,052	-,083	-,057	-,029	-,027	-,019	-,004
c1	-,064	,017	-,037	-,018	,948	,005	,003	-,003
c2	-,014	-,024	-,053	-,045	-,015	,016	-,060	-,034
c3	-,001	-,014	,007	-,016	,001	-,014	-,019	,980
c4	-,037	-,028	-,016	,979	-,016	-,041	-,009	-,017
c5	-,038	-,051	,025	-,046	,005	,909	,010	-,019
c6	,026	,000	-,004	-,009	,013	,022	,951	-,030
c7	-,158	-,184	-,201	-,192	-,117	-,087	-,067	-,049
c8	-,048	,013	,946	-,017	-,038	,019	-,004	,008
c9	,910	,011	-,050	-,034	-,065	-,025	,016	,000
c10	-,021	-,079	-,010	-,050	-,012	-,060	-,012	-,020
c11	,006	,959	,012	-,027	,013	-,046	-,010	-,015
c12	-,064	-,036	-,055	-,046	-,049	-,015	-,009	-,009
c13	-,134	-,123	,117	-,143	,004	-,312	,020	-,095
c14	-,020	,017	-,043	-,007	-,081	,062	-,008	-,003
c15	,126	-,039	-,006	-,009	,080	,017	-,024	,017
genre_rap	-,047	,061	-,045	,034	-,047	,054	,074	,093
genre_rnb	,171	-,110	,043	,113	,235	,035	,024	,069
genre_pop	-,140	-,081	-,011	,004	-,077	-,170	,316	,128
genre_hiphop	,096	-,036	-,006	,031	,030	,031	-,052	-,039
genre_dance	,234	-,171	,205	-,001	,118	-,115	-,105	-,094
genre_party	,070	-,025	,078	-,001	-,019	,102	-,002	,002
genre_soul	-,116	-,002	-,140	-,109	-,026	-,114	-,050	-,023
genre_electronic	,047	-,096	,083	,002	,083	-,114	,018	-,025
genre_rock	-,006	-,024	,034	,000	-,018	-,040	,045	,038
genre_indie	,049	-,007	-,060	,041	-,008	,113	-,028	-,019
genre_alternative	,025	-,042	-,012	,005	,010	-,003	-,052	-,036
genre_country	-,200	,041	-,066	,050	-,171	,072	-,016	-,031
genre_ballad	-,070	,115	-,060	,009	-,028	-,007	-,002	,012
genre_punk	-,066	,001	-,049	-,008	-,013	-,042	,036	,016