NATURAL LANGUAGE PROCESSING
INTELLYGENT SYSTEMS

Authors:	María Ayuso Luengo
Degree:	Master's Programme in Data Science
Date:	January 30, 2022

1 Introduction

Nowadays it's hard to spend a whole day without listening to music. Wheter it's for a personal decision, just as background or maybe it's on at the restaurant you are at music it's pretty much everywhere. Generally, songs are grouped by it's genre depending mostly on the rhythm and tempos. But are the lyrics very differentiated between genres?

In this paper we will aim to explore different songs based on the lyrics rather than genre. We will also study the predominant sentiments for each artist and decades.
The link to the code with all the results in this paper can be found at
`https://github.com/mariaayuso/Song-Sentiment-Analysis`

2 Methodology

The dataset was obtained from Kaggle ([1]) and although it keeps on being updated as people request more artists at the time of start it contained information about 21 artists such as Taylor Swift, Beyonce, ColdPlay or Ed Sheeran. Most datasets had the same format except the Ariana Grande one. We started by importing the data sets and unifying the format in order to have a single dataframe. Once we obtained it we proceeded with tha data cleaning.
After we had a general overview of the dataset, we realised some songs did not have the lyrics yet or had mistakes on other data like a Justin Bieber song that supposedely was released in the year 1. Aditionally, we realised some songs were "duplicated" since different versions of the song were included. For this reason we also removen all instances which titles contained the words "Remix", "Edit", "Acoustic" or "Live". After we had filtered these unconsistencies, we worked towards building our corpus. In this procedure we applied different word normalizetion steps. In these steps we included the transformation of the lyrics to lowercase, removal of punctuation marks as well as numbers and collapse of multiple whitespace to a single blank. In order to remove the stop words we used a collection of stop words from `tm` package and added other words that are common in songs and won't provide any information. Some examples of these are "oh", "yeah", "oww", "ey", etc. Finally we proceed with word stemming and we have our final corpus.

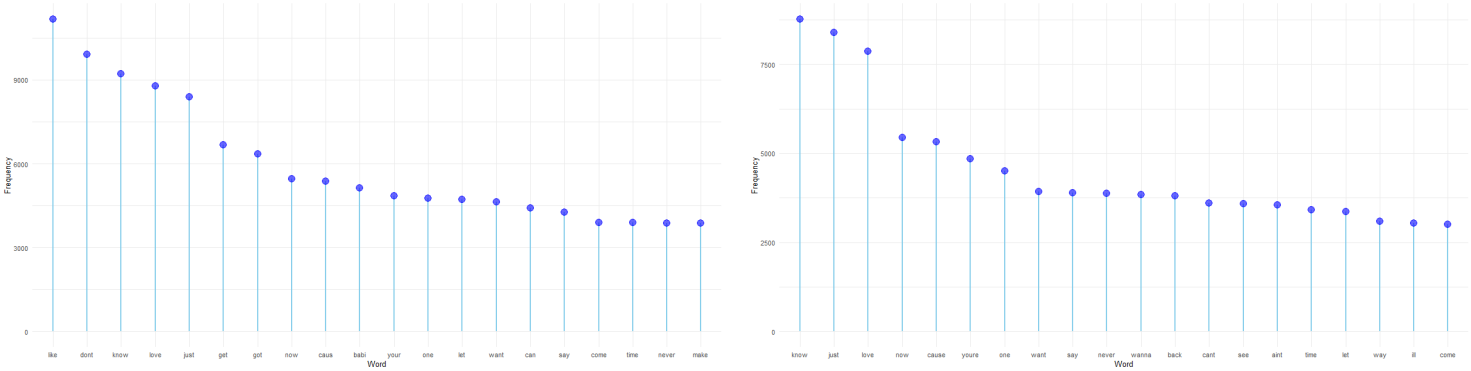
Now we will study the most frequent words and try to obtain information. As we will see in the results section this analysis wasn't very useful and we tried removing some of the most frequent words, as we could consider them stop words, to find trends in the lyrics however this wasn't very successful either. We will continue with a sentiment analysis of the lyrics. For this part, we will use the `tidytext` package since it contains sentiment lexicons that are based on single words (unigrams). We chose the Bing lexicon which categorizes words into positive and negative. Once we used this categorization we will use it to study the 20 most positive songs as well as the most negative and inspect the differences between artists.

Once the study by artist was done, we proceeded with a study grouping the dataset by decade of the song release. For this process we used the sentiment lexicon `nrc` ([2]) which not only has positive and negative sentiments but also trust, anger, anticipation, disgust, surprise, sadness, fear and joy.

3 Results

As we mentioned, once we had out corpus we started by analyzing the most frequent words. In Figure 1a we see that the most repeated words were "like", "dont" and "know". In general these and the rest of

the top 20 didn't give any information referring to the most repeated sentiments in the songs which is our final goal. On the other hand, we can see that after extending the stop words we still don't get much information since all of the words seem neutral.

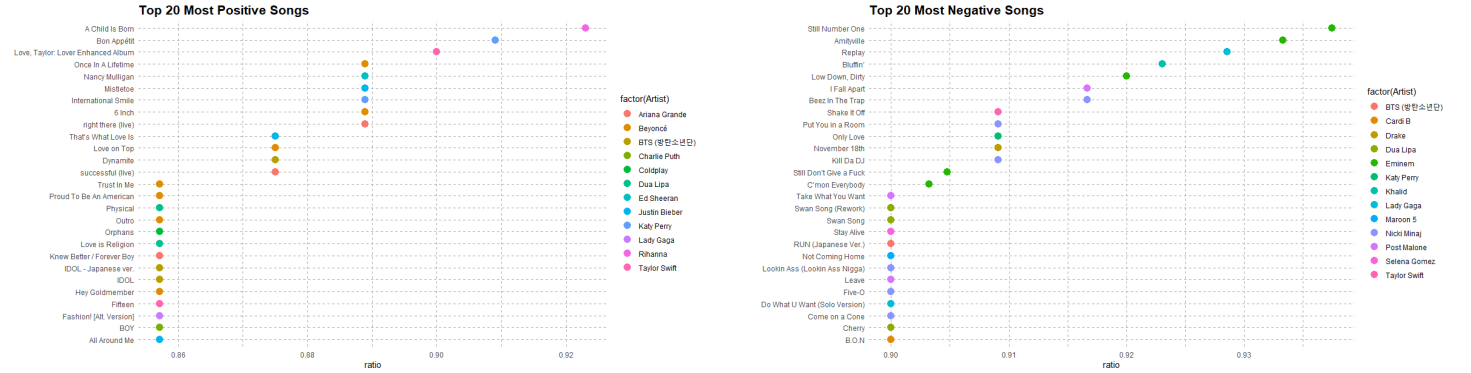


(a) Initial stopwords

(b) Extended stopwors

Figure 1: Most frequent words in corpus

Now, let's take a look at some more in depth sentiment analysis. In Figure 2a we can see the top 20 most positive songs on the dataset. First, it's important to note that as positive we mean on a relative values with the negative values of each song. Aditionally, although we used a sentiment lexicon in order to compute these ratios some positive/negative words may not appear on it and therefor the results should be validated with different lexicons. Now, going back to the 20 most positive songs we see that in general the predominant artists are Taylor Swift, Justin Bieber and Ed Sheeran. This last one is quite surprising since his hits are usually more sad or about heartbreak.
Moving on to Figure 2b these songs are mainly taken by Eminem, Lady Gaga and Nicki Minaj. On this note I think it's also important to emphasize that the words were divided only into positive or negative categories and not given a score on how positive or how negative were they. This means that all negative words are treated as the same and this can lead to misleading results.



(a) Top 20 most positive songs

(b) Top 20 most negative songs

Figure 2: Extreme sentiment songs

We will now explore the frequencies of each of the two sentiment categories for each artist. In this case, since we are looking at the absolute values it's also good to comapre it with 3 which has the total number of songs of each artist in the dataset.

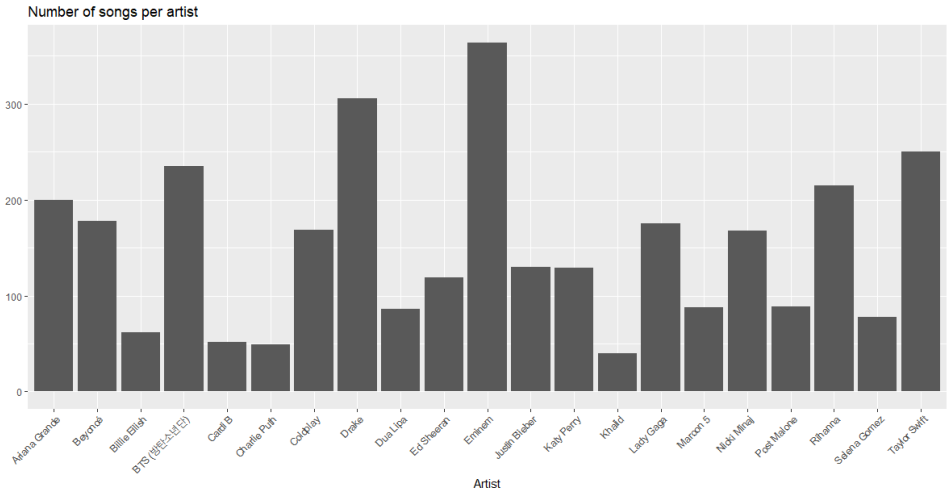


Figure 3: Number of songs of each artist

In Figure 4 we can see that generally for most artists negative words are dominant over positive. In Eminem's case this difference is the largest where the negative words are more than double the positives, whereas in the rest of them although is higher this difference isn't as remarkable. As the exception we can see that in Justin Bieber the positive exceeds slightly the negative words.

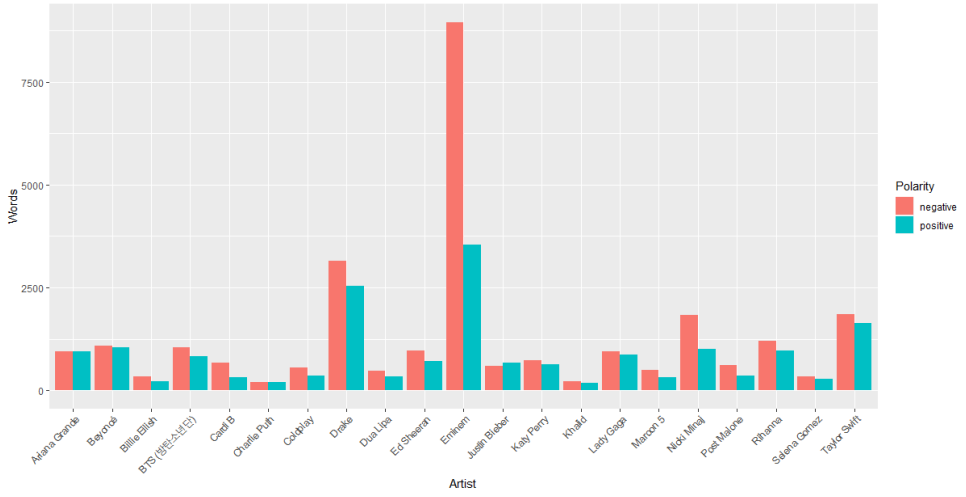


Figure 4: Positive vs negative sentiment of each artist

Looking now at the songs grouped by the decade they were released at we see different patterns. In this case, in Figure 5 we can see a clear change from the 80's to the 2000's. While both in the 80's and 90's there's a clear peak of negative words in the lyrics this starts to change by the year 2000. After this year, although the negative seems to mantain a high level it is compensated by the increase of positive words. Referring to the rest of possible sentiments, there's nothing as clear as what we pointed out although if we had to mention something joy seems to increase a little at the same time as positive which makes sense.

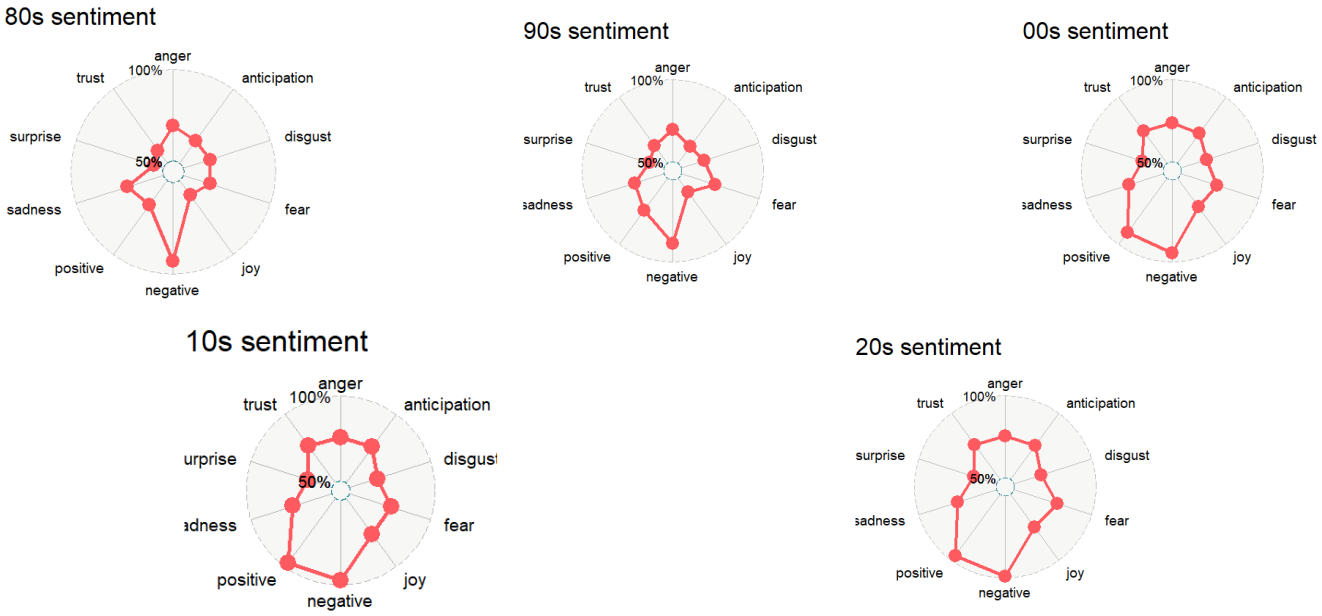


Figure 5: Sentiment analysis by decade

4 Discussion and conclusion

In this study we applied different Natural Language Processing tools to clean and create a corpus, and applied different visualization idioms to obtain information from it. Although we didn't find anything specially surprising, it's interesting to know more about the artists we here on the radio.

In terms of artist we saw that the most outstanding trend for almost every artist was negative sentiments, and also that Drake tops the rest of artists in this study. On the other hand, when looking at the release decades we saw how positive sentiment has raised since the 2000.

It would be interesting to keep on studying this domain to find clear patterns for hit songs or many other things. Further work can vary from applying a cluster algorithm to expanding the dataset.

References

[1] Deep Shah. Song lyrics dataset.

[2] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[3] Debbie Liske. Tidy sentiment analysis in r. 2018.