NLP deliverable

Problem to solve

Music is something that is part of all our lives, and the message that the artist wants to convey with the lyrics is as important as the rhythm of the song. This project will focus on the lyrics of the songs. Each song has its own feeling, with the lyrics conveying different feelings, both positive and negative. The question I want to answer in this project is whether the feeling of the songs is common in the different songs of the same artist, or how this evolves over the albums or over the years.

The lyrics of the songs I am going to analyze are from one of the most popular groups in the world nowadays, BTS. The reason for choosing this group has been personal and because their discography is very extensive, since their debut in 2013 they have released 16 albums and more than 200 songs.

Experiment(s) done

A dataset available on the Kaggle website[1] has been used to carry out this project. This dataset contains information on the songs of all the albums they have released up to date, as well as the lyrics translated into English for each of the songs, as the original language is Korean. In addition to other variables, such as the title of the song, what position it had on the album, who sings the song or if it was a remix.

Although not all the songs available in the dataset have been considered, as remixes have been eliminated in order not to have redundant information, the instrumental songs, as well as a type of "song" that appears in some albums called "skit", which is not really a song but a conversation of the members. Another change made to the dataset is the separation of the date variable, in order to have the year of the release of each song.

Once the final dataset is cleaned, the next step is to clean the column containing the lyrics, to keep only those words that are going to provide information. For this, all letters are converted to lower case, contractions are replaced (e.g., "isn't" becomes "is not"), punctuation marks and numbers are also removed. In addition, words with 2 or less letters, and words that are considered stopwords in English, are also removed. For all these operations, the "tm" library has been used.

After carrying out these operations, we have a text with those words that can be considered as providing information in the first place. But it is necessary to go one step further, to stay only with the steam of the words, to stemming.

[1] https://www.kaggle.com/kailic/bts-lyrics

The next step was to tokenise the lyrics, and to visualise which words are the most repeated throughout the songs in a general way, distributed across albums and years. The results of part of this analysis can be seen in the following section.

A sentiment analysis of song lyrics was then performed using three lexicons, "nrc", "bing" and "afinn". New datasets with the sentiment associated with each word were created for each of the sentiment lexicons. To do this, we perform an inner_join of our data containing the tokenised lyrics, along with the sentiments extracted using the "get_sentiments" function with each of the 3 sentiment lexicons discussed above.

Analysis of results

Most frequent words in lyrics

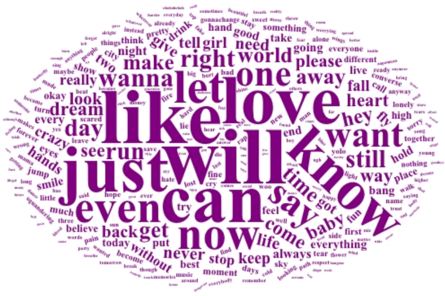

Figure 1 Most frequent words

Figure 1 shows which words are the most used in all the songs of the group. The most repeated word is will, and other words that are repeated more frequently are like, can and love.

In addition, I carried out an analysis of the most repeated words per album and per year, which can be seen in Figure 2 per album. In general, the most repeated words on each of the albums are the most repeated words in general in their discography. But it is interesting that on some albums, such as number 3 and number 13, the most repeated word is especially due to a song on that album, which I can tell from my knowledge of each of the songs.

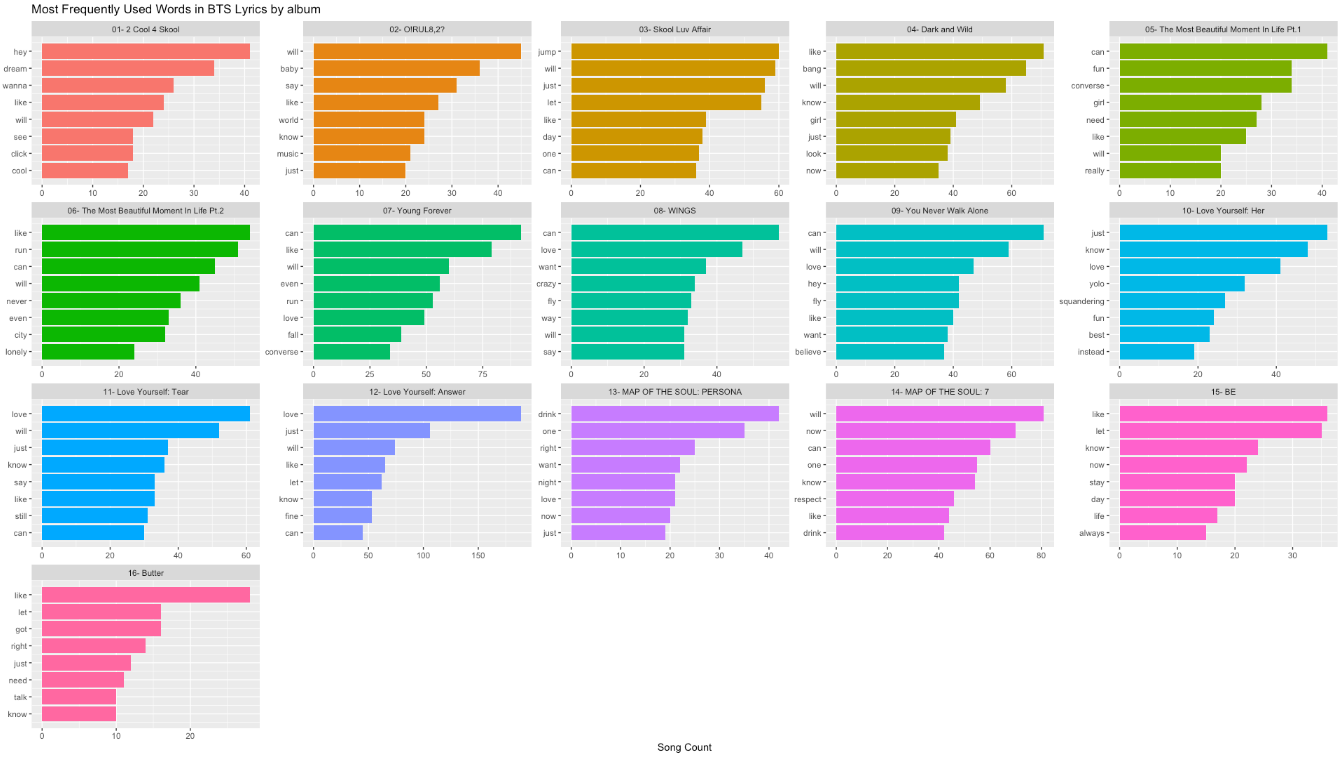

Figure 2 Most frequent words by album

Sentiment analysis

As mentioned above, a sentiment analysis has been carried out considering different lexicons, each of which provides different information. In addition, this sentiment analysis has been carried out on the discography as a whole and on each of the albums. In this case it has not been done by year, since there are several albums per year, and analyzing the albums provides more information. In this document I will discuss the most interesting results obtained, although more graphs and results have been obtained than are shown here.

First, in figure 3, we look at the results using the 'bing' lexicon, which classifies between positive and negative sentiment. In this figure we can see how these sentiments are distributed across the albums. It can be observed that in general the type of sentiment is evenly distributed. Something that can be emphasized is that in the middle of their discography the albums do have a more negative tone and it can be observed that the last album, which contains only two songs, is much more positive than the rest.

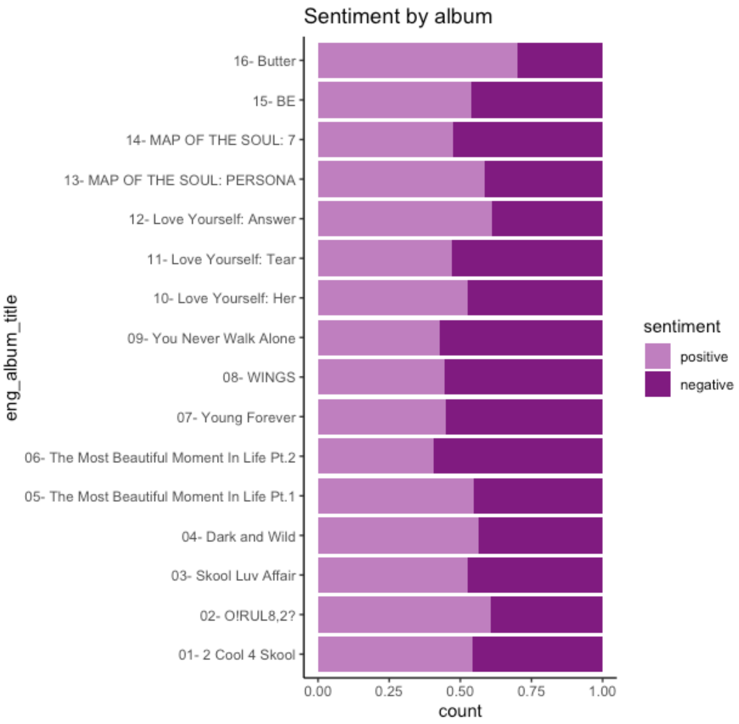

Figure 3 Sentiment by album

Sentiment analysis has also been carried out taking into account the lexicons 'NRC'. This lexicón is able to classify on different sentiments. In figure 4 we can see which are the most repeated sentiments throughout the songs. We can see that the most frequently repeated is the positive sentiment, followed by the negative, the next most repeated feelings are joy, sadness and fear, highlighting joy as opposed to the following most repeated ones. And the least repeated is surprised. This analysis provides us with additional

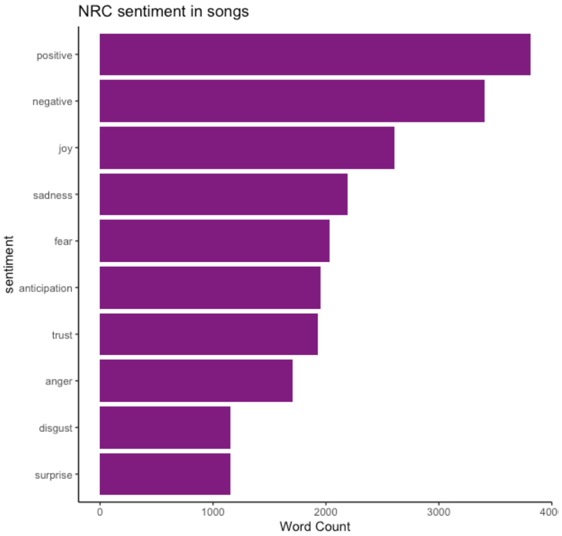

Figure 4 NRC sentiment in songs

information to that which we had obtained previously, as we can see, although they are similar values, in general the songs have a more positive than negative feeling. This can also be seen in Figure 5, where the sentiment analysis with the 'Afinn' lexicón can be found. In this case it is classified between positive or negative sentiment, but in a range of values between -5 and 5. It can be seen that there are more positive values, and that in both types of the values are found to a greater extent in the middle values.

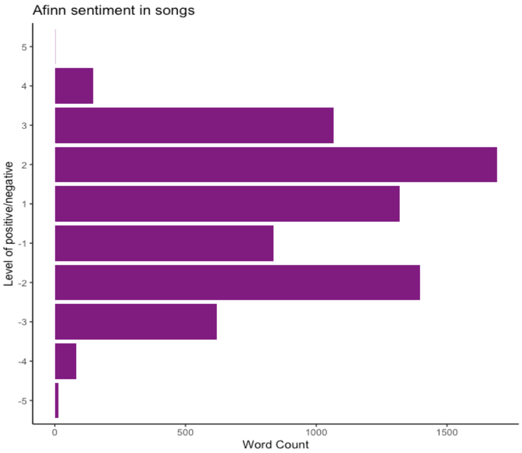

Figure 5 Afinn sentiment in songs

Using the lexicon 'NRC' an analysis of the most common sentiments in each of the albums was also carried out, as can be seen in figure 6. It can be seen that the most common sentiments per album are either positive or negative, and those where negative is the most common are consistent with the analysis previously conducted, the one that can be seen in Figure 3. It can also be observed that in the albums where the positive sentiment is the most common, usually the sentiment of joy is in second or third place. Whereas when it is negative, this occurs with the feeling of sadness.

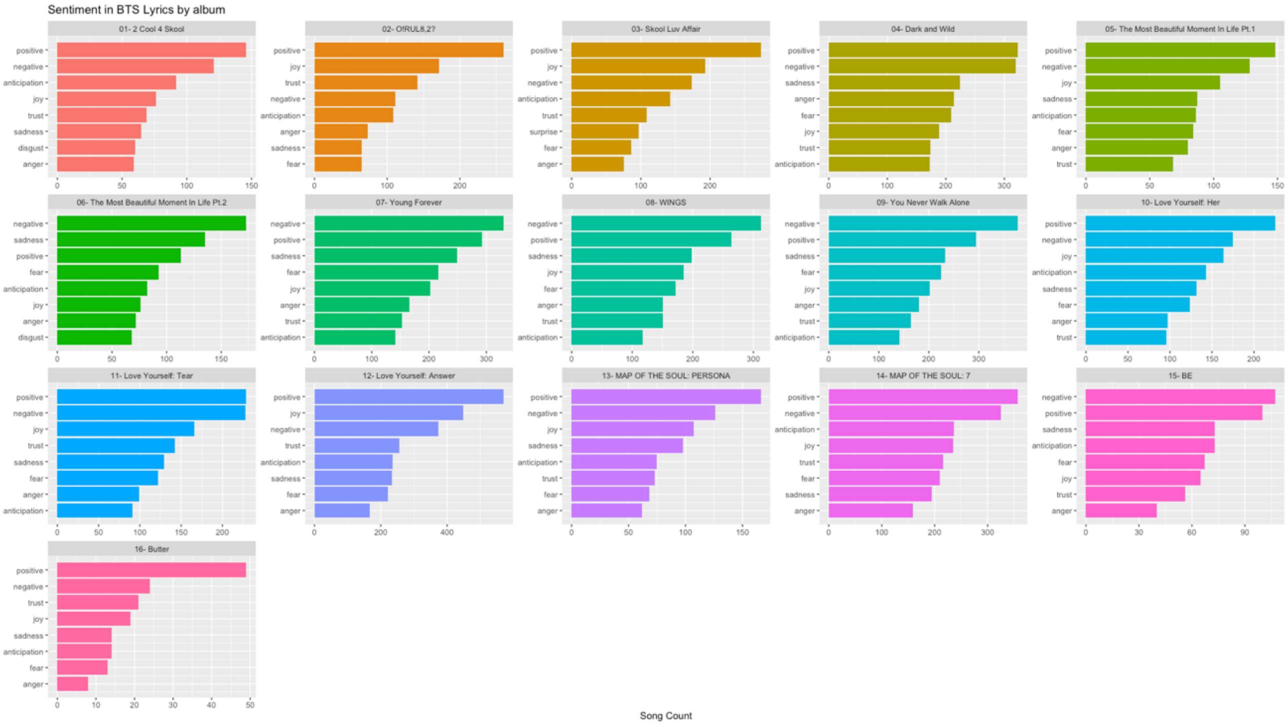

Figure 6 Sentiment analysis with NRC by album