R-bloggers

R news and tutorials contributed by (750) R bloggers

What’s in the words? Comparing artists and lyrics with R.

March 20, 2017
By Sascha W.

(This article was first published on **Rcrastinate**, and kindly contributed to R-bloggers)

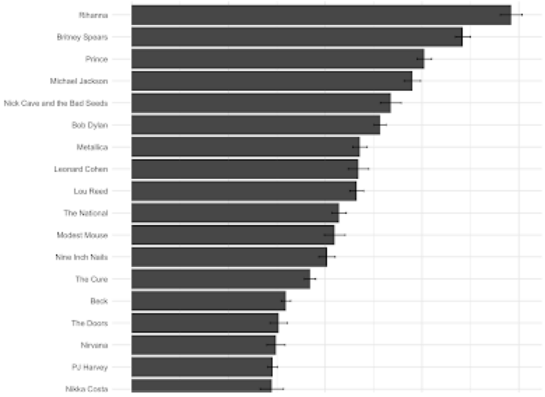
It’s been a while since I had the opportunity to post something on music. Let’s get back to that.

I got my hands on some song lyrics by a range of artists. (I have an R script to download all lyrics for a given artist from a lyrics website. Since these lyrics are protected by copyright law, I cannot share the download script here, but I can show some of the analyses I made with the lyrics.)

My main question is: What can we learn about an artist, or several artists, when we have a corpus of lyrics. I gonna analyze lyrics by the following artists:

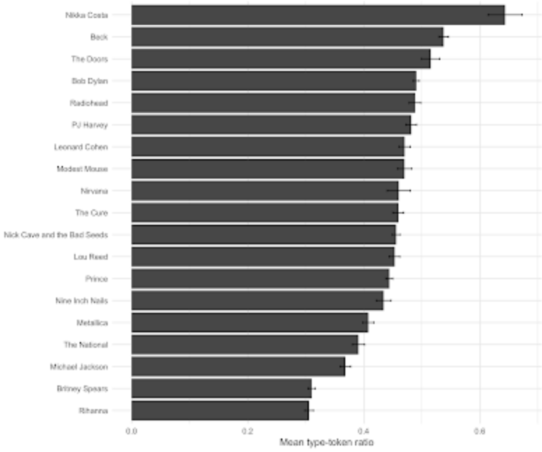
- Beck
- Bob Dylan
- Britney Spears
- Leonard Cohen
- Lou Reed
- Metallica
- Michael Jackson
- Modest Mouse
- Nick Cave and TBS
- Nikka Costa
- Nine Inch Nails
- Nirvana
- PJ Harvey
- Prince
- Radiohead
- Rihanna
- The Cure
- The Doors
- The National

Let’s start with an easy one. I wanna know which artist has the longest songs. The more words there are in the respective lyrics, the longer the song.



Mean length of songs in words (click to enlarge).

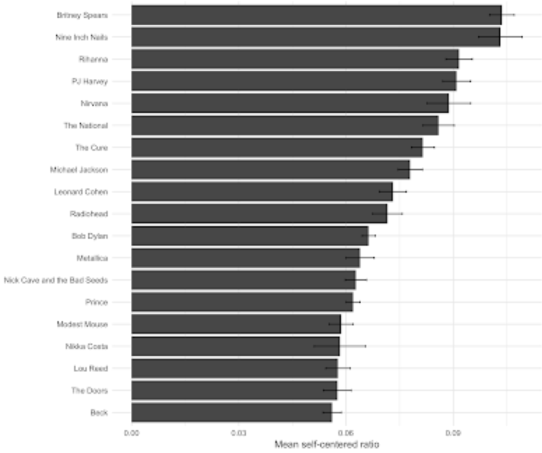
That’s quite a surprise (at least for me). Rihanna and Britney Spears, certainly the most prototypical actual pop artists in the list, have actually pretty long lyrics. Another measure from linguistics is the type-token ratio where the number of different words (types) is divided by the total number of words (tokens). This measure is often interpreted as “lexical diversity” because the vocabulary is more diverse if there are only a few words that are repeated very often. Suppose you have a song that only consists of the words “oh yeah” and this is repeated 10 times, you will have 2 types and 20 tokens, which would lead to a type-token ratio of 2/20=0.1.



Mean type-token ratio of songs (higher means more diverse vocabulary, click to enlarge).

Well, look at that – Nikka Costa, one of my favorite funk/soul artists comes out on top in this list, followed by Beck and The Doors. Rihanna and Britney obviously have a lot of words in their songs, but with regard to lexical diversity, they rank last within the artists analysed here.

Let’s try something content-related. Obviously, it’s quite hard to tackle the content (or even meaning) of songs. But we can do some really easy stuff. The first thing I want to try is what I want to call the “self-centered ratio”. I simply define a list of keywords (or better: sequences of characters) that are referencing the first person: “i”, “me”, “i’ve”, “i’m”, “my”, “mine”, “myself”. Now I calculate for each song how many of the words in the lyrics are in this list and divide this number by the number of words in the song. Suppose you have a song with these lyrics: “i’m my enemy and my enemy is mine” (I really don’t know what that would mean but that’s just an example, right?). The “self-centered ratio” would be 4/8 = 0.5 because we have “i’m”, “my”, “my” and “mine” and 8 words altogether (“i’m” is counted as one word here because it is not separated by a space). Here is the result.

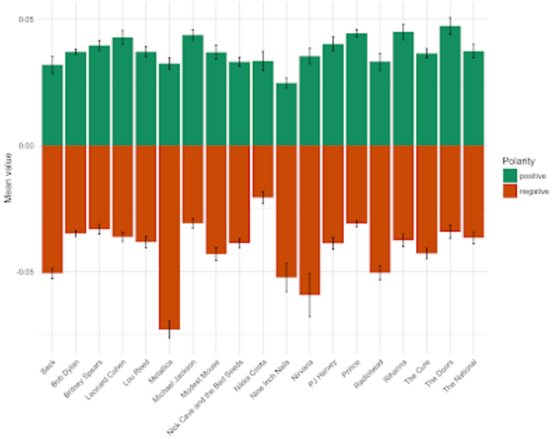


Mean self-centered ratio of songs (click to enlarge)

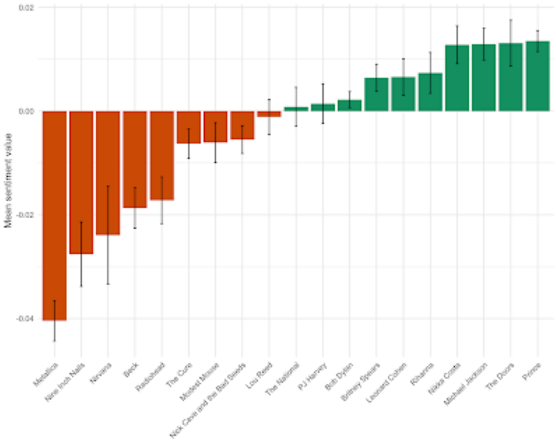
Britney and the Nine Inch Nails are definitely not very similar in terms of their music (that’s a wild guess, I only know very few songs by Britney Spears!), but they are quite similar when it comes to singing about stuff that concerns themselves.

Next up is sentiment analysis. Professionally, I don’t like it very much because in my opinion, it has a lot of empirical and methodological problems. But why not give it a try for this application here? We’re not here for the hard science side of things, are we? So, what I did was basically the same as for the self-centered ratio but only with much bigger keyword lists for positive words and negative words (so, actually I did it twice, one time with positive words and one time with negative words). I got the word lists from here (for negative words) and from here (for positive words).

I show you two plots, one where you can see both ratios and one where I combined both ratios per song to get one value (positive value + negative value). These are the results:



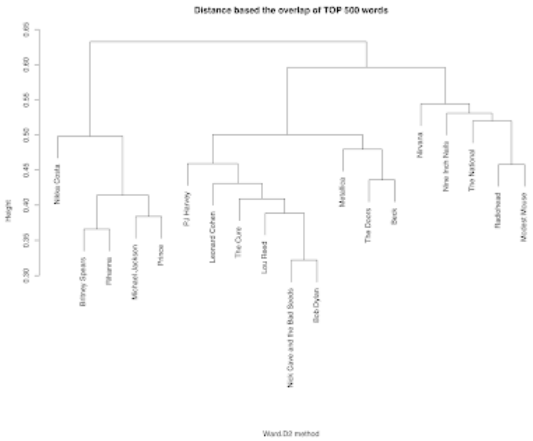
Mean ratio of positive and negative words of songs (click to enlarge).



Mean combined measure for sentiment of songs (click to enlarge).

Actually, this seems to make sense. I’m no expert for Metallica, but for Nine Inch Nails, Nirvana and Radiohead, this second plot seems to make sense. Also, Prince, Michael Jackson, Nikka Costa, Rihanna and Britney Spears getting an overall positive score works for me. Nick Cave is sometimes called the “Prince of Darkness”. In this analysis, however, this is not really confirmed. Or the “dark” aspects of his lyrics are just hidden from this quite coarse approach. Just think of the song “People just ain’t no good”. Here, each occurence of “good” is counted as positive because my simple word list approach is simply not sensitive for the negation in this line.

One last thing: I wanted to know if artists can be clustered (grouped) just with the use of their lyrics. What we need is a measure of dissimilarity for each artist-artist combination. There are several ways to do that and I experimented with a few (e.g. cosine distance or correlation of frequency vectors). It turns out, there is an even easier measure to do this: Let’s take the first 500 most frequent word each artist uses in their lyrics. With the other artist, we do the same. Then, we intersect these two sets of word lists and divide it by 500. What we get is the ratio of words that are present in both top-500 vocabularies, which is essentially a similarity measure. If we do 1 minus this value, we get a dissmilarity measure which we can use as input to a hierarchical cluster analysis. This is what we get.



Dendrogram for a hierarchical cluster analysis of overlapping top-500 words.

Look at that, I think it works quite nice: We get a “pop” cluster on the left with Nikka Costa, Britney Spears, Rihanna, Michael Jackson and Prince. Feel free to interpret the other clusters in the comments. As I said, I think it works quite OK.

R CODE is coming soon!

LOOK, there are frequency plots available here for all the artists!

Related

It's a dirty job, but someone's got to do it.. A tidytext analysis of Faith No More lyrics - Is this a midlife crisis? I wanted to ease myself back into text October 28, 2017 In "R bloggers"	It's a dirty job, but someone's got to do it.. A tidytext analysis of Faith No More lyrics - Is this a midlife crisis? I wanted to ease myself back into text October 28, 2017 In "R bloggers"	It's a dirty job, but someone's got to do it.. A tidytext analysis of Faith No More lyrics - Is this a midlife crisis? I wanted to ease myself back into text mining, October 28, 2017 In "R bloggers"

73 SHARES Share Tweet

To **leave a comment** for the author, please follow the link and comment on their blog: **Rcrastinate**.

Comments are closed.