

Comparing songlyrics from different decades with Natural Language Processing

Alina Vanessa Platz

January 2022

With the help of Natural Language Processing (NLP) I want to find differences and similarities between popular songs from different decades. I want to see how the messages music is transporting changes with time and how it reflects in cultural and political changes.

Some of the top songs from the 1960s, 1970s, 2000s and 2010s will be analysed. For this word analysis and sentiment analysis is used. A big corpus of songtexts from various artists is used for Topic Modeling. The needed data is crawled from the internet with help of the python library BeautifulSoup [2].

For conclusions connections between songtexts and important events from their respective decade were drawn. Also topics found with the latent dirichlet allocation topic model were analysed and named.

1 Introduction

Music existed for thousands of years. Why and how it evolved is unclear but it is a part of every human culture since thousands of years. The first musical instrument that was found is 40.000 years old.

Through the inventing of the radio, TV and computers not only the way of life changed but also the music and its representation began to change. Radio and TV helped with the distribution and nowadays there are endless possibilities to alter your music with the help from computers. Music was getting more experimental and old rules were challenged[3].

With the globalisation in the middle of the 20th century started the spread of popular music - music which is shared with a wide population in contrast to traditional or folk music [4]. Today there is a broad spectrum of different styles of music from around the world to which you can listen from everywhere on your phone.

Because music is and was present for a long time, I want to find out, if and how changes in our society are reflected in song lyrics. Song lyrics are widely available for free in the internet, which is perfect for crawling and analyzing them with the help of a computer. The songs analysed are english popular songs often from american artists.

2 Methodology and data

To find out how music changed the most popular songs from the 1960, the 1970, the 2000 and 2010 were crawled. To be objectice and representative they were chosen from a charts list. From the 1960s 87 songs were crawled, from the 1970s 86 songs, from the 2000s 82 songs and from the 2010s 84 songs.

To have a enough data for topic modeling songs from some of the top artists (see appendix for whole list) from 1970-1979 were crawled.

2.1 Scrapping the lyrics

For scrapping the python library BeautifulSoup [2] was used. Most URLs from the site [1] from which I crawled the lyrics are build like this:

"/lyrics/<artist>/<song>.html".

For every decade a list with the top hundred songs [8] is copied. A function reads in the artist and the song and adds them to the URL above. When the URL exists the lyrics will be crawled.

For the bigger corpus the top artists from the 70s were given to the programm. The programm uses the name of each artist and their first letter to create an URL like this:

"/[a-z]/<artist>.html"

From this site all URLs that link to a song are isolated. Those URLs were used to crawl the songtexts from the artist.

To avoid being blocked from access to the site, after every scrap the function waits for five to fifteen seconds, before scrapping the next songtext.

2.2 Topic Modeling with Latent Dirichlet Allocation

In Topic Modeling three terms are defined:

1. Word
This is the smallest entity. Each word is represented as a distinct vector.
2. Document
A document consists of a sequeze of words.
3. Corpus
A corpus is a collection of multiple documents.

Topics are a group of words which appear often together in a document. For Topic Modeling the Latent Dirichlet Allocation (LDA) algorithm was used. LDA is generative probabilistic model of a corpus.

For topic modeling the songtexts from the 1970s were used. After deleting all songtexts that were not recognized as english by langdetect [5] the data contains 5064 songlyrics from 27 artists.

In the first step the data was cleaned. To do this all characters that are not letters were removed from the documents. Music has a structure with repeating elements that do not transport additional informations. Those duplicate lines



Figure 2: Words only used in 1960-1969 (top left), 1970 - 1979 (top right), 2000-2009 (bottom left) and 2010-2019 (bottom right)

The most used words didn't changed much (see figure 3 in the appendix for the wordcloud from each decade). More interesting was how the use of words changed over time.

Next I searched for words that were just used in one decade. From 2000-2020 there were more new words in songs used, than words that were only used from 1960 to 1980 (see figure 2).

A word that was only used in 1970 - 1979 is "watergate". Watergate is used to describe a bunch of political scandals between 1972 and 1974. Those scandals amplified even more the existing mistrust against politics [9]. The word water-gat is used in two songs.

The first one is "Sweet Home Alabama" from Lynyrd Skynyrd.

"Now Watergate does not bother me
Does your conscience bother you?"

The second song is "Bicycle Race" from Queen.

"I don't wanna be a candidate
For Vietnam or Watergate
'Cause all I want to do is
Bicycle bicycle bicycle"

Those critical mentions are hidden in happy and innocent sounding songs and capture the feeling of disenchantment with politics in this time. Also it captures the essence of the popular hippie phrase "Make love not war" which

was also an important cultural part of this decade.

A word that was replaced and used only from 1960-1979 is "highway". In 2000-2019 only the word "freeway" was used. Those two words are not completely the same as every freeway is a highway but not every highway is a freeway [10].

The context in which highway appears is different than the context in which freeway appears. Whereas highway are mostly connected with adventures and speed, a freeway in the newer songs seems to be more slow and relaxed.

"Get your motor runnin' Head out on the highway Lookin' for adventure" - Born to be wild, Steppenwolf	"Hold me closer, tiny dancer Count the headlights on the highway Lay me down in sheets of linen" - Tiny dancer, Elton John
"I'm on the highway to hell" - Highway to hell, AC/DC	"You and me girl got a highway to the sky" - Too much heaven, Bee Gees

"Every day's just a holiday Cruisin' on the freeway" - Just lose it, Eminem	"Skin that flick, she's such a little DJ To get there quick by street but not the freeway" - By the way, Red Hot Chili Peppers
---	--

In reality a freeway is less likely to have speed limits [10]. An explanation for this could be that the highways were massively expanded starting in 1956. More than 42.000 miles of interstate highways were built [11]. Maybe because of the novelty of broad access to highways it was perceived as more interesting and adventurous.

The language used 2000-2019 is more vulgar and aggressive than in the decades before. There are many words about partys, drug use and crimes. But there are also words like "haunted" and "madness".

The words capture an excess and a feeling of being lost.

In the 1960s the word "party" appeared in one song and in the 1970s party didn't even appear once. In the 2000s "party" appeared in five songs and in the

2010s in two songs. Additionally appeared the verb of party "partying" the first time in two songs in the 2010s.

3.2 Similies

Songtexts usually use figures of speech. One figure of speech are similies. Similies are easy to spot with dependency matching because they are always build like this:

1. as [verb] as [noun]
2. [verb] like [noun]

Some similies found in the songtexts are:

1. follow like sheep
2. burning like flame
3. seemed like cardboard
4. swim like dolphins
5. shines like sun
6. shakes like fish
7. fly like planes
8. feel like circus
9. stick like glue
10. feel like home
11. floating like butterfly
12. stinging like bee

3.3 Topics in songs

With a LDA Topic Model eight topics were found. Those topics are "love", "faith", "crime", "friendship", "loneliness", "flirting", "party/dancing" and "sadness" (see Table 1 for the top terms of the topics).

The topic that appeared the most was "love". The least used topics were "friendship" and "faith".

Top Terms	Associated Topic
Love, know, time, heart, want, way, feel, let, tell, life, hold	Love
la, hear, sing, lord, come, time, people, song, god, world	Faith
bad, going, get, got, like, come, right, ready, well, know	Crimes
little, man, boy, like, play, old, life, mother, song, sing, change, brother	Friendship
away, day, go, come, time, run, look, night, like, home, leave , find	Loneliness
get, know, say, good, tell, man, go, way, hey, time, girl, women, come	Flirting
baby, come, want, get, let, wanna, girl, to, love, know, rock	Partys and dancing
lie, eye, light, cold, night, burn, die, high, soul, fire, cry, life, tear	Sadness

Table 1: Words that appeared together in documnts and their associated topic

4 Conclusion

Music transports feelings of their decade and are a reflection of the emotions of their time. However the deeper meanings are often hidden and computers alone can not grasp them. Especially for topic modeling this turned out to be a obstacle, because words are often used out of context as a metaphor.

Computers are extremly helpful in filtering interesting aspects which then have to analysed further by a human. With more historical knowledge there are more connections one can make.

For further exploration of this topic there should be more songs for every decade be choosen to analyse. One can find interesting connections by just using a small amount but to be more precise there should be more music to be analysed.

Another interesting aspect is if and how the topics changed overtime. For example is the topic "faith" still appearing as much in the 2000s as it appeared in the 70s? Are there new topics that emerged with new technologies?

Out of 5064 songs 1167 could with a precision less than 40% assigned to one of these topics. This problem couldn't be solved with more topics. An explanation for this is that songs can have more than one topic and that in songs words are often used in figures of speech and not as literal. This is a shortcoming that should be considered for the future.

References

- [1] Az lyrics. <https://www.azlyrics.com/>.

- [2] Beautiful soup. <https://beautiful-soup-4.readthedocs.io/en/latest/>.
- [3] History of music. https://en.wikipedia.org/wiki/History_of_music.
- [4] Popular music. https://en.wikipedia.org/wiki/Popular_music.
- [5] Project description langdetect. <https://pypi.org/project/langdetect/>.
- [6] Spacy dependency matcher api. <https://spacy.io/api/dependencymatcher>.
- [7] Spacy phrase matcher api. <https://spacy.io/api/phrasematcher>.
- [8] Top 100 songs. <http://www.top-100-songs.de/index.html>.
- [9] Watergate-affäre. <https://de.wikipedia.org/wiki/Watergate-Affäre>.
- [10] Jane Lu. Freeway vs. highway: What's the difference? <https://getjerry.com/insights/freeway-vs-highway-difference>.
- [11] Matt Rosenberg. How wwii created the interstate highways. <https://www.thoughtco.com/interstate-highways-1435785>.

A Lists

Used Artists from the 70s

- ABBA
- AC/DC
- Aerosmith
- Al Green
- Barry White
- Bee Gees
- Black Sabbath
- Blue Swede
- Boston
- Carpenters
- Deep Purple

- Demis Roussos
- Dolly Parton
- Eric Clapton
- Five Stairsteps
- George Harrison
- Gilbert O’Sullivan
- Hollies
- Iggy Pop
- Led Zeppelin
- Marmalade
- Nazareth
- Norman Greenbaum
- Pink Floyd
- Queen
- Rolling Stones
- Smokie

B Figures



Figure 3: Most used words in 1960-1969 (top left), 1970-1979 (top right), 2000-2009 (bottom left) and 2010-2019 (bottom right)