Modern times

This is an example of the high art of international diplomacy. The great subtlety and cultural finesse of the american president helps building the delicate relationships with foreign parties.

Wait, what !?

We have online witch hunts, because someone used the wrong word in the wrong context.
And yet, over 600k people actually „liked" this post.
It seems fair to assume that not all of them do it for the comedian value (not everyone likes dark humor).

This is not an isolated incident, actually it seems symptomatic. So we wonder:
Are angry people online actually concerned—or is this just group dynamics ?

Can we find data to investigate this?

We need uncensored text. Ideally we want to relate it to popularity or social acceptance. And we also want to take a look at the development over time ...

Analysing Offensiveness in Music

Data

We use the 1 Million Song Dataset, which offers us a list of songs with corresponding metainformation, such as popularity as well as lyrics in a bag-of-words format. Here's what we did:

Quantifying offensiveness (as a non-native speaker):

The British telecommunications provider OFCOM compiled a list of words ranked by their offensiveness, from „mild" to „strongest".
We scanned the lyrics for occurences of these words and then checked every match manually.
Some offensive words have been removed since they are only offensive in context. The bag of words approach only allows us to see individual wordcounts, no complete phrases. „Jesus Christ" for example was filtered out.

Quantifying popularity

We use a „hotness" field from the metadata of the sonds. Unfortunately this is missing for many entries.
So we also use information on playcounts from the Nest streaming service. The idea is that a song that was played often, is probably popular.

Conclusions ?

We were not able to find any significant dependencies between popularity, offensiveness and time.
Of course, in data science no answer is also an answer.
But it feels like there should be something.
Call it research bias, but we decided to revisit the data.

Quantifying offensiveness ?

The dataset doesn't contain the original lyrics, because of copyright. Unfortunately, the bag-of-words approach limits our analysis to single offensive words. However, we do think that the our offensiveness rating is precise enough.

Quantifying popularity ?

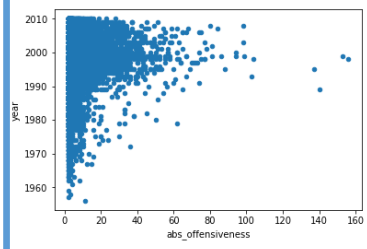
The hotness attribute is missing for many songs. We intended to base our analysis on the playcounts. But a comparison of the recorded playcounts with numbers on video views from youtube show that the data is not trustworthy. Very popular songs have been played only a few times, according to our dataset.
To continue our research, we would first invest time into a better metric for popularity

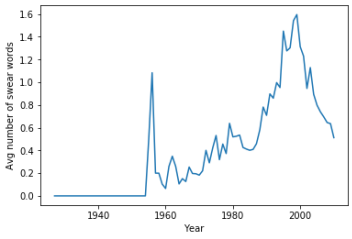
Insights

Offensiveness over time

The Scatterplot seems to have some structure

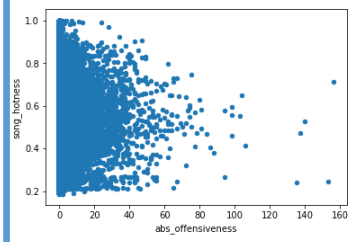
But there is no trend in the average number of swearwords





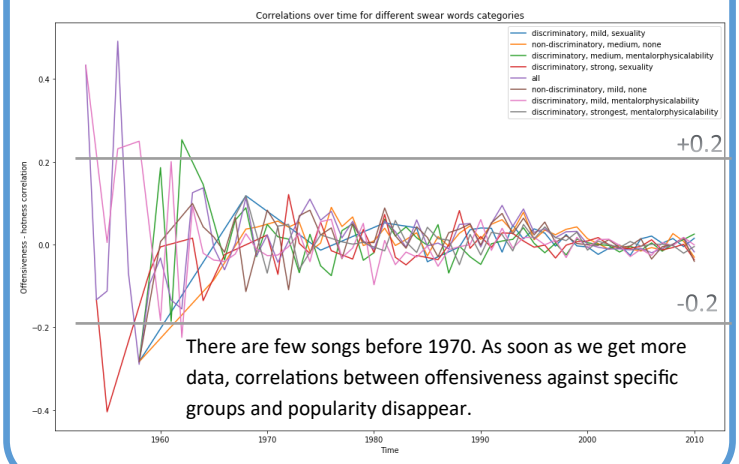
We also looked at the percentage of offensive songs over time: How many songs that have been released in a year would be classified as offensive. We tried different metrics and threshold for this classification. But there doesn't seem a significant dependency on time.

What about popularity ?



At a first glance, it seems like songs that are either very popular or very unpopular or not offensive. Offensive songs are usually just „average".

Are we missing a more complex relationship ? We calculated correlations based on the different target groups, over time.



There are few songs before 1970. As soon as we get more data, correlations between offensiveness against specific groups and popularity disappear.

"What is freedom of expression? Without the freedom to offend, it ceases to exist."
—Salman Rushdie