# The Evolution of German Rap Lyrics

Luca Schumann
*lucsc442*
*TDDE16*

***Abstract*—Test**

## I. Introduction

Introduce the task or research question that you have addressed in your project. What were you trying to do? Why did you choose this project?

## II. Theory

Present relevant theoretical background, and in particular the models that you have used. Where appropriate, use mathematical formulas.

### A. Latent Dirichlet Allocation

### B. Principal Component Analysis

## III. Data

In this Section I present the source and selection criteria for the data as well as the preprocessing steps for the songs.

### A. Dataset

First of all, a selection of songs had to be made. To find a balance between relevant artists in the CD era and the streaming era, I decided to combine lists of artists from two sources. I selected all 46 german artists from the Wikipedia list of Rappers that sold more than 100,000 albums in Germany [1]. The popularity of musicians today, especially Rappers, is not only defined by the albums sales but also the online streams. Therefore I picked all 13 Rappers appearing in the Spotify playlist "Top German Artists of 2020" [2] that were not in the selection yet. To increase the dataset, I chose 18 more objectively popular Rappers, adding up to 77 artists in the end.
The lyrics of the artists' songs were then downloaded using the Genius API[1] with help of the *lyricsgenius* package [3]. I excluded feature songs, as it would have been very difficult to seperate which artist sung which part. A lot of the songs are duplicates, since remixes, features and live versions are all included in the genius database. There are furthermore many non-songs in the database, such as commentaries. Luckily, most could easily be ignored by discarding songs that contain brackets, as in *Papa ist da (AsadJohn Remix)*. Few others had to be handled differently, but the final dataset is clean of such occurences. In a first run of the model, songs containing english lyrics formed an individual topic. Upon manual inspection, there were a few songs with parts in english, or even completely in english. Therefore, songs that contain all the most common english words in the english topic "the", "you" and "and" were suspected to contain english lyrics and also discarded. The final dataset features almost 7500 songs.

[1]https://docs.genius.com/

### B. Preprocessing

Preprocessing the song lyrics was an iterative process with many learned lessons. The most basic first step is to remove meta information that is written in brackets from the lyrics, such as *[Hook]*. Next, the songs are tokenized using spacy's German package. Non-alphabetic tokens, stop words and tokens with less than three letters are removed. The rest of the preprocessing is done with Gensim. Bigrams and Trigrams are added and extremes are filtered. Finding the seemingly best values for these tasks was done with the trial and error method.
After excluding english songs from the dataset, the model would still find a topic with mostly english stop words. After removing english stop words during preprocessing, the problem was solved. Another problem was many artist's names appearing in the most salient words of some topics. The reason for this is that some Rappers are notorious for talking and singing about themselves in the third person. Rappers with a lot of songs in the dataset would therefore skew the results, although their own name has nothing to do with the message behind their song. To tackle the problem, each artists name was removed from his or her lyrics. Unfortunately this method is not foolproof, as some refer to themselves by nicknames, e.g. *Ufo361* simply calling himself *Ufo*. As a final tweak of the preprocessing step, a couple of stop words that were not in spacy's German list were manually removed.
Tokenizing German text turned out to be a difficult task, with no really satisfying solution. In German texts, all nouns are capitalized, therefore also in spacy's word list. If a word at the beginning of a sentence is capitalized, it is not possible for spacy to tell wether it's a noun or not. Feeding spacy all words in lowercase would also not solve the problem, since it would not recognize the nouns anymore. Additionally, some words have a different meaning depending on wether they are a noun (capitalized) or not. E.g. "Weg" means path but "weg" means away in German. This appears to be an unsolved problem as of today, as can be read on spacy's Github forum [4]. Lemmatization does also not work as well as for english text, but overall the results are good enough to work with them.

## IV. Method

All the code for the project can be found inside the jupyter notebook in my Github repository[2].

First the lyrics of the selected artists are downloaded from the genius API with help of the lyricsgenius package [3] and stored in individual json files. Next, the lyrics are extracted from the json files and some first preprocessing is done: Non-songs, songs with English parts and duplicate songs are dropped, linebreaks and meta information are removed from the lyrics. The remainder is then stored in a data frame together with the release date, artist name and song name.

Next, the lyrics are lemmatized using spacy's German word list. Stop words, non-alphabetic tokens and short tokens are additionally removed during this step. Afterwards some manually added stop words, english stop words and Rapper's names are removed. The reasoning behind this is explained in Section III-B. The rest of the preprocessing is done using gensim. Bigrams and trigrams are added, the tokens are added to the dictionary, tokens that appear to often or too seldom are removed from the dictionary and finally the bag of words representation is computed to be stored in the corpus.

Now the LDA model is computed for five topics and other parameters that can be found in the notebook. After the model is finished, the topic distribution for each song is computed and added to the data frame. Two more columns are added, one with the majority label of each song and another one with the popularity score from Spotify[3], downloaded using the spotipy package [5]. The final data frame is also uploaded to the repository as *all_songs.pkl*, so to reproduce the results the data does not need to be redownloaded. The plots shown in this report are all created from that data frame, the code for them is also in the notebook.

## V. Results

The results are split into four parts, the first being about the discovered topics and the others about further analysis of the dataset.

### A. Model and Topics

To find the best model and number of topics, the coherence scores for a Non-negative matrix factorization (NMF) and a LDA model with 5-15 topics were computed. LDA with five topics is the only model that achieves a score higher than 0.4. Previous work has shown [6] that manual quality assessment is better suited for topic modelling song lyrics, which is why I additionally manually chose the best model/topic number combination. As a result, it is also the best choice from my personal point of view, as the resulting topic clustering delivers the greatest expressiveness compared to the other options.

The final iteration of the model found five topics that I named "Street", "Sex & Party", "Love & Life", "Competition" and "Lifestyle". To show the limit of the classification, I manually labeled one song of every artist with the expected topic. The model managed to find the expected label in 57 of 77 songs (74%). Table I shows the eleven most salient terms for each topic, translated to English. It strucks right away that many English terms are used in the lyrics, especially in the "Lifestyle" topic, were seven words are not translated. Interestingly, there are none in "Street" or "Love & Life". Some words such as *see*, *man* or *yeah* make it in the top terms of multiple categories. *Come* for example even appears twice very high up in the "Sex & Party" list. In the original German list the verb is once capitalized, which is the reason for it appearing twice. As mentioned in Section III-B, spacy does not recognize capitalized non-nouns at the beginning of sentences correctly.

TABLE I
11 Most Salient Terms for each Topic, Translated

| Street | Sex & Party | Love & Life | Competition | Lifestyle |
|---|---|---|---|---|
| brother | baby$^{O}$ | life | Rapper$^{O}$ | bro$^{S}$ |
| money | come | (to) love | Rap$^{O}$ | bitch$^{O}$ |
| street | know | world | see | money$^{O}$ |
| mom | come | see | come | money |
| head | yeah$^{O}$ | know | boy | bitches$^{O}$ |
| boys | say | stay | yeah$^{O}$ | yeah$^{O}$ |
| out | please | stand | bitch$^{O}$ | cash$^{O}$ |
| life | party$^{O}$ | man | fuck | gang$^{O}$ |
| block | drag$^{M}$ | think | man | bro$^{S}$ |
| away | night | just$^{M}$ | king$^{O}$ | fuck |
| run | club$^{O}$ | heart | people | Gucci$^{O}$ |

$^{O}$Original, not translated $^{M}$Multiple meanings possible $^{S}$Slang

Figure 1 shows the intertopic distance map, visualized by the Python implementation of the LDAvis tool [7]. The five topics are drawn as circles in 2D space, by first computing the intertopic distances and then multiplying those with the transformation matrix obtained from PCA. Each topic's overall frequency is depicted by the size of the respective blob. The topics "Lifestyle" and "Sex & Party" stand out as the smallest, with 11.8% and 10%, respectively. They also have a huge gap between them and the other topics. The other three topics are rather close in comparison, but still with a noticeable distance between each other. The next largest topic is "Street" with 15.4%, followed by "Competition" with 29.1% and finally "Love & Life", which is assigned to 33.7% of all tokens.

In Figure 2 you can see the intertopic distance map for a model trained to use six topics. This serves as a comparison to the distance map of the final model. The two smallest blobs are assigned 4.4% and 5.3% of all tokens, respectively. The topics 1, 2 and 3 are arranged similarly as in Figure 1, albeit with a shorter distance inbetween. Additionally, there is topic 4, which is very close to topics 1 and 2, indicating a major overlap with both topics.

### B. Topics in Rap over Time

In this section I visualize the development of German Hip Hop lyrics over the last 30 years. Figure 3 depicts the average topic distribution of all songs that were released in the respective year. Since the number of songs released differs

[2]https://github.com/Lucapaulo/Text-Mining-Project

[3]https://developer.spotify.com/documentation/web-api/

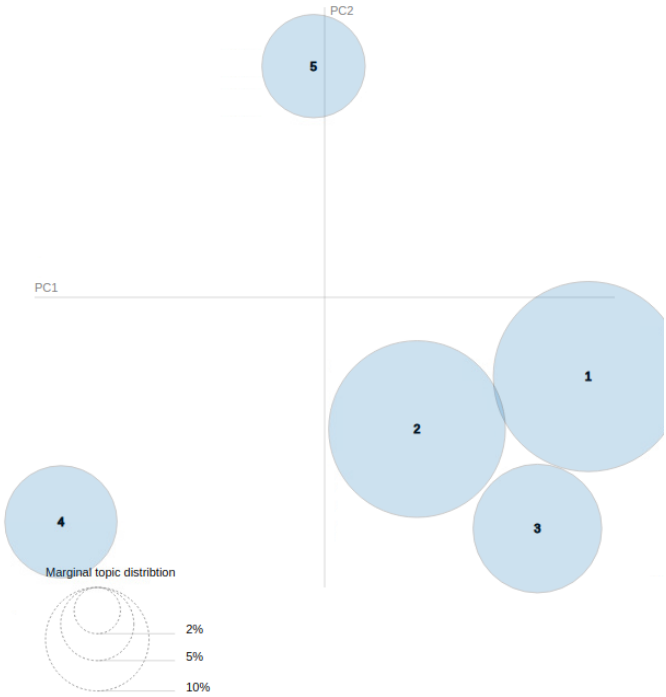


Fig. 1. The intertopic distance map for all five topics. The distance is computed as the Jensen-Shannon divergence and projected to 2D space. 1: Love & Life, 2: Competition, 3: Street, 4: Lifestyle, 5: Sex & Party.

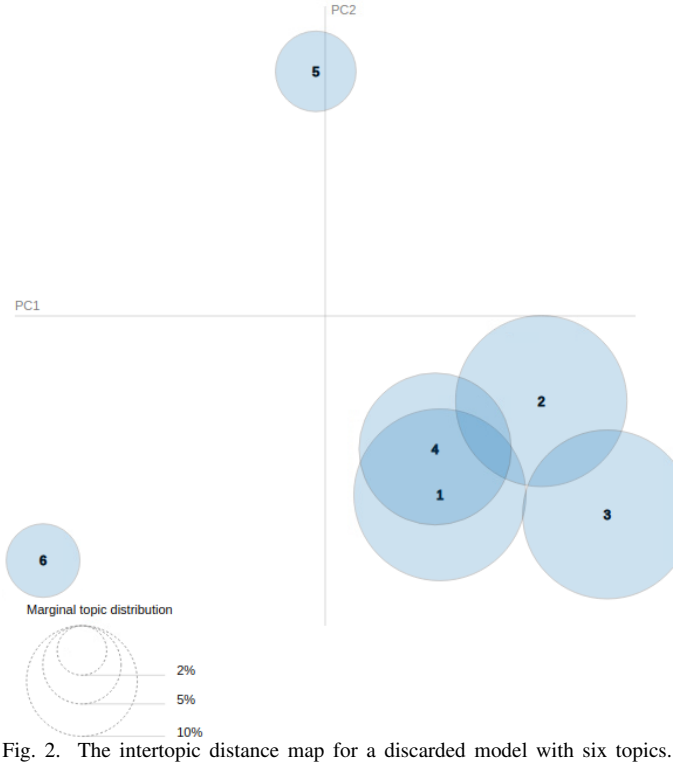


Fig. 2. The intertopic distance map for a discarded model with six topics. The distance is computed as the Jensen-Shannon divergence and projected to 2D space.

each year, the data is normalized to better capture the bigger picture. Note that they years until 2003 contain less than 100 songs each, whereas each year after 2011 has more than 400

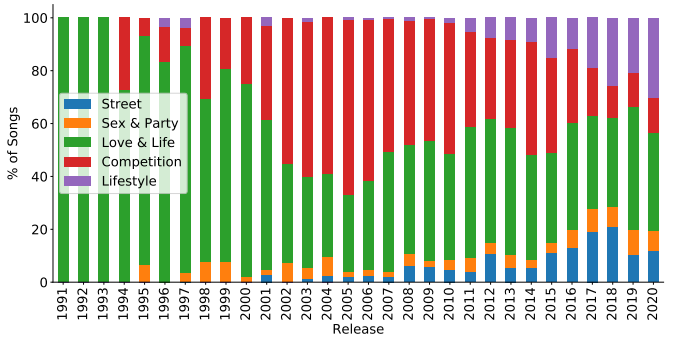


Fig. 3. Normalized timeline of the average topic distribution in the lyrics of songs released between 1991 and 2020.

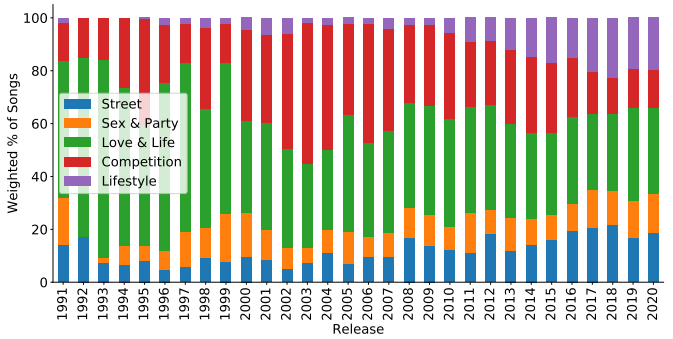


Fig. 4. Normalized timeline of the average topic distribution in the lyrics of songs released between 1991 and 2020. Songs are weighted by their popularity score on Spotify.

songs in the dataset. This imbalance is not due to a poorly select subset of artists, but rather due to the rapid growth in popularity of German Rap in the last 10 years [8]. Multiple trends can be seen in this graphic. First of all, an increase of the topic ”Competition” with a peak in 2005, followed by a steady decline until 2018. Next, we observe a fast growth of the ”Lifestyle” topic since 2010, as well as a moderate growth of ”Street” Rap between 2008 and 2018. Throughout most years, ”Love & Life” appears to be the most dominant topic.

This method automatically gives higher influence to artists with more songs in a release year. It also disregards the music preferences of the audience. Figure 4 is meant to capture the preferred topics by the consumers, by weighting each song with its Spotify popularity score. The score is downloaded for each song from the Spotify API using the *spotipy* package [5]. It measures the current popularity of a song on a scale from 0 to 100. To increase the influence of more popular songs on the graphic, I extended the scale to 1000. An exponential function maps e.g. the score 100 to 1000 or 90 to ca. 500. Compared to the unweighted graphic, we immediately notice a much higher share of the topics ”Street” and ”Sex & Party” throughout the entire timeline. The topics ”Love & Life” and ”Competition” on the other hand seem to be less popular. ”Lifestyle” is mostly unchanged by the weighting.

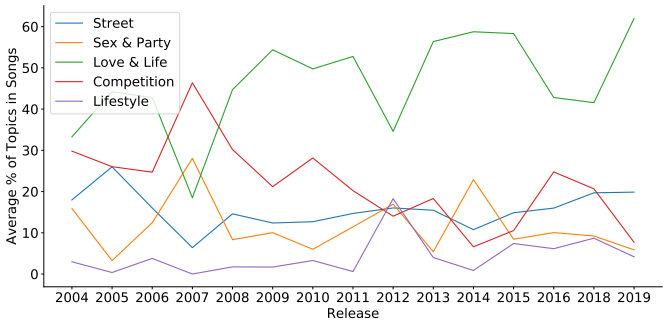


Fig. 5. The average percentage of topics in *Sido*'s songs between 2004 and 2019.

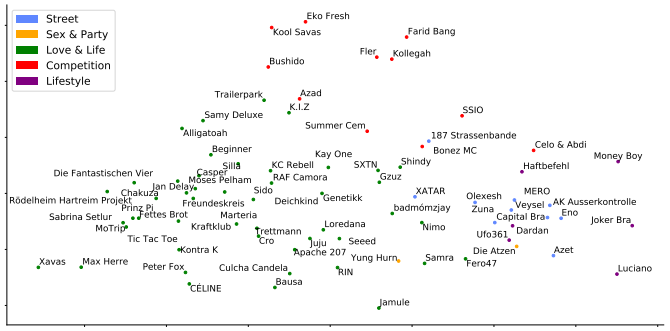


Fig. 6. A 2D visualization of the average topic distribution in each artist's lyrics. The dots are colored by the majority label.

### C. Sido's Story - Underground to Mainstream

In this section I want to lay the focus on one particular Rapper, namely *Sido*. Although he started his music career in the late 90s, his solo career did not start before the end of 2003, which is why the graph in Figure 5 starts in 2004. 2004 to 2007 we can see an upwards trend of the topics "Competition" and "Sex & Party", contributing to almost 80% of the Rapper's lyrics in 2007. The following years, there is a decline of the two topics, which are being contested by "Love & Life". 2019 "Love & Life" makes up around 60% of the lyrics on its own. "Street" appears to always play a role in his songs, hovering around 15% throughout the years. "Lifestyle" has the lowest contribution to his music in each year, apart from a 20% peak in 2012.

### D. An Artist Overview

The last part of this section focusses on the similarity of the lyrics of all analyzed artists. Figure 6 visualizes the distance between the artists in 2D space. The average distribution of the artists' songs for each of the five topics were the original vectors in 5D space, reduced to two dimensions using PCA. The dots are colored with the majority label of each artist's lyrics. We can see a large "Love & Life" cluster on the left side of the graph, that makes up more than half of the entire space. The upper part is dominated by the "Competition" topic, and the right side is shared by "Street" and "Lifestyle". "Sex & Party" as the majority label does not show up in an individual cluster, it only features two artists that are rather separated.

## VI. Discussion

### A. Analysis

In this subsection I will give an interpretation to the previously shown results of section V.

*1) The Topics:* The choice for the correct number of topics is of most critical importance for a good topic model. I finally sticked to five topics, as it gave the best balance between expressiveness and quality of results. Of course a larger topic number improves the ability of the model to correctly assign a topic to a song. On the other hand, clogging the graphs with 20 different topics makes them harder to interpret, so instead I decided to keep the topic number small and the topics broader. The sizes of each topic vary to a reasonable extent, so that no topic is underrepresented. The intertopic distances are also large enough, such that there is not too much overlap between them, but a relatively clear distinction. Other models have shown far worse results, with extremely niche/broad topics or heavily overlapping topics, as shown on the basis of the six topic example in Figure 2. On a sidenote, NFM seemed to always be slightly worse than LDA, which is why I sticked to LDA.

Table I is meant to help understand the content of the topics better. Some of these words seem not to fit the topic on the first read, because they may have a specific meaning in Rap culture. E.g. when *king* is mentioned in Rap lyrics, the meaning is supposedly more often *king of Rap* than that an actual Monarch is meant. The difference in the number of English and slang words between the topics can potentially mean that language has an influence on the topic choice. Two sentences with the same meaning but a different word choice could be assigned to different topics. The word *yeah* could have been added to the stop word list, but it is hard to decide wether it really belongs there.

*2) Rap over Time:* The rise of the "Competition" topic in both Figures 3 and 4 strongly correlates with interesting developments in the German Rap-Scene, with the first rivalries that were carried out in public via Battle-Rap songs. The first big "beef" was between Azad and Samy Deluxe in 2001. Azad provoked Samy Deluxe in his song "Gegen den Strom" (against the stream), who then reacted with his disstrack "Rache ist s" (revenge is sweet), which then again was countered by Azad with "Samy De Bitch". This is just one of countless battles that took place [9], with a peak after 2004, where famous Rapper Bushido left his label Aggro Berlin with help from the criminal Abou-Chaker clan [10].

"Lifestyle" is a topic mainly represented by artists of the Trap genre, which started to establish in Germany with artists like Money Boy, RIN, Yung Hurn and Ufo361. The growing popularity of this subgenre is also underlined by the fact, that each of the just named artists can be found in Spotify's "Top Artists of 2020" playlist [2].

Streetrap has existed for a long time in Germany, as can be seen in 4, but has been through a revival that started with Haftbefehl in 2009. This revival phase was amplified by many now prominent Street Rappers like Bonez MC,

Gzuz and RAF Camora [11]. Figure 3 also shows, that the volume of songs representing this style has only increased since then. From the popularity chart we can take away that the listening preferences of Spotify consumers show a strong bias towards the topics "Street" and "Sex & Party". Not only recent songs, but also older ones seem to be more popular nowadays compared to contemporaneously produced songs.

*3) Sido:* The radical change in the lyrics depicted in Figure 5 correlates with *Sido*'s development as an artist. Until the start of his solo career, he was wearing a chromium skull mask and known as a Gangsta-Rapper. His first album as a solo artist was furthermore heavily criticised and indexed [12] by the *Federal Review Board for Media Harmful to Minors* [13] because of misogynous and drug-glorifying texts. Today, the Rapper is regularly represented in the Charts with rather "harmless" songs, often featuring Pop artists [14].

*4) Overview:* The largest part of Rappers in Figure 6 being labeled as mainly "Love & Life" is no surprise. First of all it is the most prevalent topic in the model. Secondly, the topic contains many common words, such as (translated) to know, to see or to say. It is noticeable, that almost all of the artists that were active in the 90s in the set are inside the green cluster, most even very far on the left. Some examples for that are *Fettes Brot*, *Tic Tac Toe*, *Rödelheim Hartreim Projekt* and *Die Fantastischen Vier*.

The top side of the graph features Rappers like *Fler*, *Bushido* and *Kollegah* that are often involved in Rap battles [9]. Most of them also have a career going for 15 years or longer, which means they were active during the "Competition" peak found in Section VI-A2. This red cluster is a strong indication that the "Competition" topic is correctly extracted.

The only representatives of "Sex & Party" as the majority label here are *Yung Hurn* and *Die Atzen*. Both sing almost exclusively about drugs in *Yung Hurn's* case or about party in *Die Atzen's* case, which also makes their labels reasonable.

The artists *Gzuz* and *Bonez MC* are very close to each other as well as to their group project *187 Strassenbande*, despite all three having a different majority label. This can be interpreted as an even distribution of three topics in the three artists' songs.

Finally, the right side of the graph features many active street Rappers [11] such as *Capital Bra*, *Mero* and *Ufo361*. Purple colored dots are also spread around in the same area, which points towards a strong correlation of the topics "Street" and "Lifestyle".

### B. Limitations

With more time and work the preprocessing could be further improved by tackling the already mentioned problems. Smaller influences on the results like removing Rappers' nicknames but also bigger ones like improving the quality of stemming. The different language styles of Rappers are also a limitation, but since I use the bag of words representation of the words, the model should be able to understand the context of those words to an extent. Of course the dataset could also be bigger. One option would be to include feature songs and seperate the parts of each artist for the analysis of the individual topic preferences. Another one would be to simply increase the number of Rappers in the dataset. Time is mostly the limiting factor here, since downloading lyrics from the genius API is very slow and aborts regularly, resulting in many restarts.

The small number of topics is another limiting factor, as they do obviously not cover everything and are too broad in some cases. As explained in Section VI-A1, this choice was made to have expressive Figures to show. For a more in depth analysis of lyrics the number of topics could be increased in future work.

### C. Related Work

There has already been a lot of previous work in the topic of analyzing song lyrics. In a similar work, Sasaki et al. [15] used LDA to extract the weights of five latent topics for each song in their data. The authors used Japanese songs for their model, my work extends theirs to German lyrics.

Another alike approach from Tsukuda et al. [16] proposes a novel method to model song lyrics which incorporates the artist's taste for topics.

Kleedorfer et al. [17] use NMF to identify topic clusters in mostly English lyrics. I found out that NMF performed worse in my experiments, which is why I used LDA instead.

Sterckx et al. [6] use the kurtois metric to assess the quality of a topic model for lyrics, as it strongly correlates with manual topic quality scores. They find that existing semantic measures like the coherence score are not very suitable for this task, which is why I decide to manually find the best topics.

Work on German song lyrics is very limited. Roman Schneider [18] created a corpus of German lyrics together with linguistic features and extralinguistic metadata. His work is different in content and exclusively focuses on Pop songs, wheares this project analyzes songs of the Rap genre.

## VII. Conclusion

Based on your results and their analysis, what new knowledge do you take away from your project?

## References


[1] Wikipedia. (2020) Liste der meistverkauften Rapalben in Deutschland. [Online]. Available: https://de.wikipedia.org/wiki/Liste_der_meistverkauften_Rapalben_in_Deutschland

[2] Spotify. (2020) Top Künstler*innen 2020. [Online]. Available: https://open.spotify.com/playlist/37i9dQZF1DWTdV9tXbHOAv

[3] J. W. Miller. LyricsGenius: a Python client for the genius.com API. [Online]. Available: https://github.com/johnwmillr/LyricsGenius

[4] Github. (2018) Improve rule-based lemmatization and replace lookups. [Online]. Available: https://github.com/explosion/spaCy/issues/2668

[5] P. Lamere. Welcome to Spotipy! [Online]. Available: https://spotipy.readthedocs.io/en/2.16.1/

[6] L. Sterckx, T. Demeester, J. Deleu, L. Mertens, and C. Develder, "Assessing quality of unsupervised topics in song lyrics," in *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, ser. Lecture Notes in Computer Science, M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, Eds., vol. 8416. Springer, 2014, pp. 547–552. [Online]. Available: https://doi.org/10.1007/978-3-319-06028-6_55

[7] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, jun 2014, pp. 63–70. [Online]. Available: https://www.aclweb.org/anthology/W14-3110

[8] G. S. Dr. Florian Drücke, Sigrid Herrenbrück, "Musikindustrie in Zahlen 2019," 2020.

[9] Wikipedia. Liste von disstracks des deutschen hip-hops. [Online]. Available: https://de.wikipedia.org/wiki/Liste_von_Disstracks_des_deutschen_Hip-Hops

[10] J. Schaaf, "Bushido vs. arafat abou-chaker." [Online]. Available: https://www.faz.net/aktuell/gesellschaft/kriminalitaet/bushido-vs-abou-chaker-ich-war-auf-dem-beifahrersitz-ganz-eingeschuechtert-16931630.html

[11] O. Marquart. Eine dekade im rückblick #3: Straenrap neue deutsche welle.

[12] NRWE. (2006) Verwaltungsgericht Köln, 27 K 6557/05. [Online]. Available: http://www.justiz.nrw.de/nrwe/ovgs/vg_koeln/j2006/27_K_6557_05urteil20060217.html

[13] "Tasks and responsibilities of the federal review board for media harmful to minors." [Online]. Available: https://www.bundespruefstelle.de/bpjm/meta/en

[14] Billboard. Chart history sido. [Online]. Available: https://www.billboard.com/music/sido/chart-history/GES

[15] S. Sasaki, K. Yoshii, T. Nakano, M. Goto, and S. Morishima, "Lyricsradar: A lyrics retrieval system based on latent topics of lyrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, H. Wang, Y. Yang, and J. H. Lee, Eds., 2014, pp. 585–590. [Online]. Available: http://www.terasoft.com.tw/conf/ismir2014/proceedings/T105_352_Paper.pdf

[16] K. Tsukuda, K. Ishida, and M. Goto, "Lyric jumper: A lyrics-based music exploratory web service by modeling lyrics generative process," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 544–551. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/96_Paper.pdf

[17] F. Kleedorfer, P. Knees, and T. Pohle, "Oh oh oh whoah! towards automatic topic detection in song lyrics," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 287–292. [Online]. Available: http://ismir2008.ismir.net/papers/ISMIR2008_211.pdf

[18] R. Schneider, "A corpus linguistic perspective on contemporary german pop lyrics with the multi-layer annotated "songkorpus"," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, 2020, pp. 842–848. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.105/