Detecting Text Reuse and Similarities between Artists in Rap Music through Visualization

Christofer Meinecke*, Stefan Jänicke†

*Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Leipzig, Germany
E-mail: cmeinecke@informatik.uni-leipzig.de

†IMADA, University of Southern Denmark, Odense, Denmark
E-mail: stjaenicke@imada.sdu.dk

***Abstract*—Detecting references and similarities in music lyrics can be a difficult task. Crowdsourced knowledge platforms like Genius [1] can help in this process through user annotated information about the artist and the song but fail to include visualizations to help users finding similarities and structures on a higher and more abstract level. We propose a prototype to detect and visualize the similarity of rap artists based on their lyrics and monolingual alignments of song lyrics. For this, we apply word and sentence embeddings to lyrics we crawled from Genius.**

***Index Terms*—Text Reuse, Intertextuality, Visualization**

I. Introduction

Rap music emerged from a long history and tradition as a rhetoric of resistance [2] into a standalone music genre. As of now, the hip hop music industry is one of the biggest music industries in the US but also in Germany; it is the most streamed music on platforms like Spotify [3]. Rap music as a part of the hip hop culture combines the "creative use of language and rhetorical styles and strategies" [2]. This characteristic of rap music creates similarities to literature in regards to using poetic language or referencing other artists like the rephrasing of famous quotes or from a musical stand-point through sampling. Especially the intertextuality can enhance the enjoyment of the music through emotions such as nostalgia. Nevertheless, detecting all references can be a difficult task because the listener needs a lot of knowledge of the genre and its history. The user crafted annotations from platforms like Genius [1] can help in this process. We combine different natural language processing techniques with visualizations to communicate similarities in lyrics to domain experts and casual users that are interested in music. These similarities can give starting points to further search for cases beyond rephrasing like plagiarism or can just increase the knowledge of the user about the genre and its history. Through platforms like YouTube [4], Spotify [5], or SoundCloud [6] access to new music is eased with almost no barriers. Unfortunately, this also allows to easily copy the lyrics or other characteristics of a song. This can be hard to detect especially for not well-known songs that are not written in English or the dominant languages of the country the plagiarist lives. Visualizations can be applied to communicating such similarities and to further ease the process of detecting them. In particular, the domain problem of detecting similar lines can be seen as a text alignment problem [7]. We visualize these alignments in a hierarchical way starting with an edge in a graph as an aggregate over two artists, followed by streamlines representing the songs and showing dependencies between them and finally the side-by-side inspection of two lyrics in a collation manner.

We propose a method based on word and sentence embeddings to detect similarities between rap artists based on their lyrics and to detect text reuse in their lyrics. For this, we crawl the lyrics from Genius.com [1], which are enriched with metadata about songs and the artists and additional annotations about the lyrics. Furthermore, this approach is generalizable to the lyrics of all genres of music.

II. Related Works

A. Similarity of Musicians

Similarity Analysis of Musicians is one of the applied use cases in the STAR about Visualizations for Musical Data by Khulusi et al. [8]. Although the text of a song is not directly musical data it is still connected to the music and the musician.

Widely applied methods to measure the similarity of musicians include using crowdsourced data like Amazon sale statistics [9] or Spotify listing histories [10]. Similar works are done by Gibney [11], Cano and Koppenberger [12] and Gleich [13]. All these methods visualize, similar to us, the data through node-link diagrams either focusing on a given artist or the whole database but they do not include additional visualizations to inspect a lower detail level of the data.

Similar to the former works, platforms like Genius are crowdsourced and include rich annotated metadata about musicians and more importantly the transcribed lyrics of the artist. These text collections can be analyzed in terms of text reuse and overall similarity. Some works compared the vocabulary of rap artists extracted from Genius for American [14] and German artists [15]. Another work used the vocabulary to define the similarity between the artists [16], which is also in the focus of our work.

Another way would be to observe the influence of the musicians of the past on the currently active musicians [17], [18]. These works often focus on rock music and visualize the data through graphs. Similar approaches can be of interest for rap music because of the long-existing culture of referencing and collaboration, where new upcoming artists are referencing previous artists or are supported by established artists.

In contrast to the prior works Jänicke et al. [19] designed a visual analytics system that supports the profiling of musicians based purely on biographical characteristics and excluding their works.

Some music information retrieval works focused on lyrics to compute similarities between artists but without visualizing them [20], [21].

B. Song Similarity

In contrast to our approach most music information retrieval systems focus on sound features but often combine them with the lyrics [22]. Yu et al. [23] combine textual and audio features by deep cross-modal learning to retrieve lyrics based on audio and audio based on lyrics but did not include visualization. The LyricsRadar [24] allows users to browse song lyrics while visualizing their topics in a two-dimensional vector space. Furthermore, graph-based visualizations to tackle plagiarism detection based on sound features are designed by Ono et al. [25] and De Prisco et al. [26].

C. Text Alignment

Our focus lies on textual data and has similarities to works based on textual alignment and more species textual plagiarism detection and text reuse. Text alignment application scenarios can be divided into three areas [7], first collation, which examines and records similarities and differences among variant text editions, second the detection of text reuse, like fragments, allusions or paraphrases, third translation alignments where cross-lingual connections are focused.

Common methods to visualize text reuse patterns are Grid based [27], [28], Sequence-aligned [29] or Text-oriented [30] Heat Maps. More popular are side-by-side views supported by stream graphs and aligned barcodes [31], [32]. On a line-level variant graphs [33], [34] and tabular views [35] can help to visualize similarities and differences. A detailed overview of text alignment visualizations can be found in the survey by Yousef and Jänicke [7]. From a text alignment perspective, we visualize text reuse scenarios on the song- and line-level with collation methods, where we treat similar lyrics as textual variations [32]. For this, we apply side-by-side views and variant graphs.

III. Data

Genius.com is a website where casual users and even artists themselves can transcribe lyrics of songs and annotate them with additional information. This information can include references to other songs or artists, an explanation of specific words or phrases e.g. slang or wordplay, or connections to historic or current events. Genius started as "Rap Genius" in 2009 but changed the name in 2014 to include knowledge for other music genres and other types of media like literature. Through the Genius API, the data about a specific artist, song, or annotation can be extracted, including metadata about other social media platforms, relationships to other songs, involved artists e.g. feature guests, producers, and more. Annotations can be added by every user but they need to be accepted and reviewed by a moderator. For our prototype, we crawled data from Genius for around 600 German rap artists and groups.

IV. Methodology

A. Tasks and Design Rationales

We propose a system based on word and sentence embeddings to compute the similarity between rap artists based on their lyrics collected from Genius [1]. The system refers to users from the general public interested in music and especially interested in German rap music that wants to further explore the connections between different artists, find similar lyrics or similar lines. Following Munzner's guidelines for task abstraction [36], the domain-specific tasks are to *explore* a network of musicians, to *compare* their lyrics, to *identify* similar musicians and songs based on their lyrics (points of interest to further search for potential ghostwriting, or plagiarism cases), and so to *derive* references between songs. For the line-level tasks, we applied a line-level alignment approach based on side-by-side views to allow the comparison of lyrics. For the exploration of the network and to identify similar artists we used a force-directed layout where the edges between the nodes are based on the similarity of the lyrics.

B. Artist Similarity

To compute similarity values between the artists based on their lyrics, we applied the fastText word vectors trained on the German Wikipedia corpus [37]. The benefit of fastText is in this case, to include out-of-vocabulary words, which are a common phenom for rap lyrics because of slang, adlibs or neologism. We treat each line in the lyrics as a sentence, for which a sentence vector is computed through unsupervised smooth inverse frequency [38]. Therefore, the sentence vector is a weighted average of the word vectors. The weight depends on the word frequency, the vocabulary size and the average sentence length of the corpus. The sentence vectors are added to a faiss [39] index structure to query the nearest neighbors based on cosine similarity. We focused on lines instead of sentences because rap artists write their lyrics line by line, also lines are often sentences.

The similarity s_{ab} of an artist a to an artist b is computed based on the cosine similarity of the nearest neighbors:

$$s_{ab} = \sum_{i}^{n} c_s(l_i, l_r) \cdot ((k+1) - r)$$

For two artists, we use all their lines that are nearest neighbors. For such a pair l_i and l_r, we take the cosine similarity $c_s(l_i, l_r)$ multiplied by the number of nearest neighbors $k+1$ minus the rank r of the neighbor. Through this, we get a rank-based weighting. We then take the sum over all such pairs for two artists. This value can be further normalized by using the total number of lines of all songs of the artist.

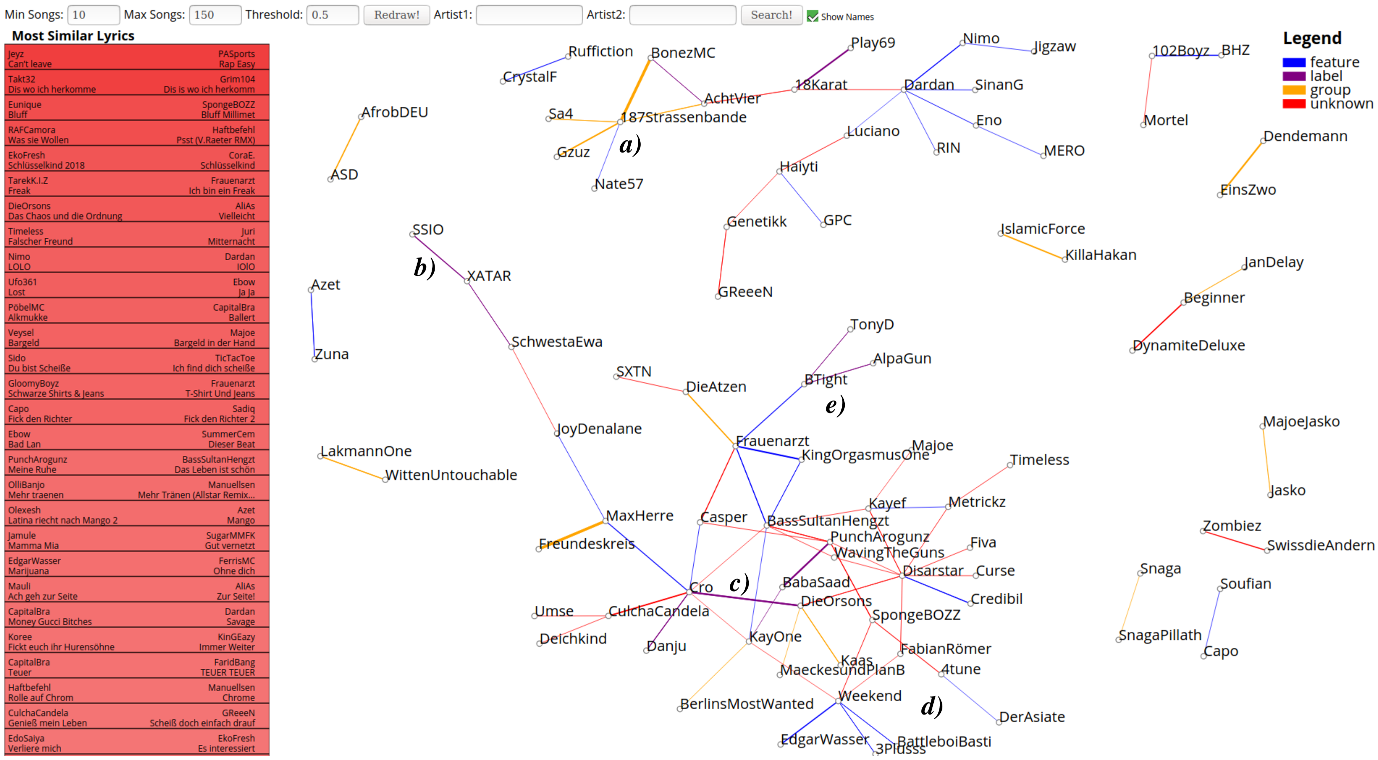

Fig. 1. An excerpt of the similarity network of German rap artists based of the most similar lines in their lyrics. Label and collaboration partners tend to be connected with each other.

C. Artist Similarity Graph

Following the Information Seeking Mantra [40], we started with giving an overview by visualizing the similarity as a graph with a force-directed layout. An edge indicates that the artists are similarly based on their lyrics. To reduce visual clutter the user can filter based on the similarity values and the minimum and maximum number of songs of an artist. After filtering, all nodes without an edge satisfying the condition are removed. Furthermore, we color-coded the edges to show different relations. A blue edge indicates that two artists have at least one song together, a purple edge indicates that the artists are or were signed by the same label, an orange edge is a “part of” relation for group members, while a red edge shows an unknown relation. We choose red for the unknown relations to better highlight them as they represent an unknown or missing social relation. The different relation types show social connections beyond the lyrics which can give hints on why the lyrics of two artists are similar. We extracted the relation types from the lyrics and the Genius metadata. Furthermore, we mapped the similarity value of two artists on the edge thickness, to highlight relations with a higher similarity value. A sub-graph can be seen in Figure 1 showing different clusters. The cluster at a) shows the 187 Stassenbande a rap group from Hamburg and another Hamburg based rapper Nate57, while b) shows a group of the Bonn based label “Alles oder Nix Records” lead by the rapper XATAR. At c) a group of the Stuttgart based label “Chimperator” with the rappers Cro, Danju and Die Orsons (Kaas, Maeckes und Plan B) can be seen. Although Danju and Cro left the label in 2017 and 2018. d) shows a cluster of rappers known for their Video Battle Tournament History like Weekend, 3Plusss, BattleboiBasti, 4tune, Der Asiate, SpongeBOZZ and PunchArogunz. The cluster at e) shows a group of Berlin based rappers like Frauenarzt, King Orgasmus One, Bass Sultan Hengzt, Die Atzen, SXTN, B-Tight, Alpa Gun and Tony D.

Next to the graph, a list of the most similar song pairs is displayed. The song pairs are color-coded from white to red on a linear scale depending on the number of nearest neighbors. When clicking on a song pair in the list the side-by-side alignment view is displayed. Furthermore, a user can search for two specific artists of interest or click on an edge in the graph to investigate the songs of the artists.

D. Artist View

An interesting property of the Genius data is the rich annotated metadata including references and information about the artists. We give an overview of some of the metadata from Genius and display a list of the most similar artists based on their lyrics and all of the songs of the artists in the artist profile view. This view is accessed when clicking on a node in the graph or the artist’s name in the side-by-side view. An example can be seen in Figure 2. The list of most similar artists is color-coded in the same way as the graph but instead of the edge thickness, saturation is used. Through this list, the user can further explore other artists. The list of songs includes for each song the ten nearest neighbors color-coded in the same way as the list of the most similar songs. Furthermore,

Fig. 2. The artist profile of the German rap group BHZ.

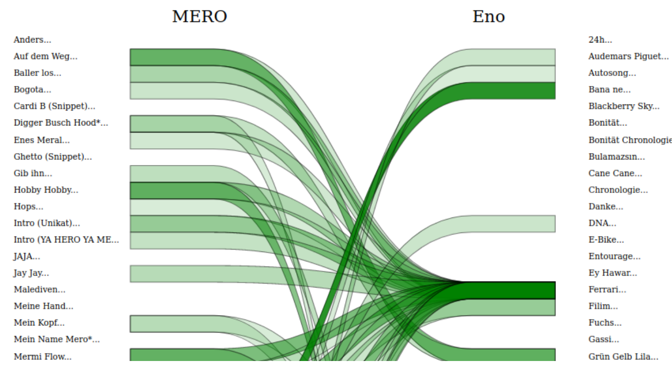

Fig. 3. Excerpt of the Song-Level of Mero and Eno.

the metadata from Genius is used to display relations with other songs. These relation types are: samples, sampled in, interpolates, interpolated by, cover of, covered by, remix of, remixed by, live version of and performed live as. By clicking on a color-coded nearest neighbor the alignment view pops up. Therefore a user can explore the network and find different points of interest to further investigate the alignments.

E. Monolingual Alignments

When investigating two artists of interest a stream graph visualizes the amount of nearest neighbors between the songs of the artists. The number of nearest neighbors is mapped on the edge saturation. An example can be seen in Figure 3. To reduce the visual clutter filter mechanism can be applied. A user can filter based on the number of nearest neighbors and the release date of the songs. The lyrics of the songs can be read when clicking on a streamline. Both song lyrics are placed side-by-side while the nearest neighbors of each line are shown similar to the visual analytics system iteal [32]. This allows a user to read the lyrics side-by-side while investigating the alignments. Each alignment is visualized as a streamline connecting the lyrics. Furthermore, the user can filter the alignments based on a slider. The filter values correspond to the cosine similarity between the lines in the alignment. This allows to further investigate the nearest neighbors of two songs of interest. Multiple examples can be seen in Figure 4. Figure 4 a) shows "Kool Savas - Komm mit mir" and "Alligatoah - Komm mit uns" where Alligatoah parodies the original song by Kool Savas. Figure 4 b) shows "Fatoni - D.I.E.T.E.R." and "Samy Deluxe - Dieter" where both artists criticize the German musician Dieter Bohlen. The excerpt in Figure 4 c) shows "Sido - Du bist Scheiße" and "Tic Tac Toe - Ich find dich scheiße" where the song by Sido sampled the original by Tic Tac Toe. When clicking on a streamline of interest the alignment is visualized as a variant graph using TraViz [33]. Furthermore, all of the nearest neighbors of both lines are shown with TraViz, as seen in Figure 5. These nearest neighbors can be used to move to another song pair of interest where the alignment occurred.

F. User Feedback

We did user studies with 5 fans of rap music that have general and scene-specific knowledge about the German rap scene. They used the system for approximately half an hour up to one hour to explore the graph and the relations between the artists and the lyrics. One user suggested adding filtering by year for the song-level side-by-side view to focus on specific parts of the artist's career e.g. when two artists were part of the same group, or if only early works or new works are similar. For example, he noticed a higher similarity in the lyrics of Tony D and Sido when both were part of the rap group "Die Sekte". A user noted that the list of similar songs in the profile view is helpful to detect songs about the same or similar topics e.g. love, cars or drugs. Another user noticed that sometimes the alignments are only created because of one common word. The reason for this is probably our approach of including all of the ten nearest neighbors of each line and also the usage of the cosine similarity. So including a threshold and other metrics could be helpful. Often alignments are between the hook or the refrain of two songs so for future works it would be better to treat them in a different way to focus more on less clear similarities.

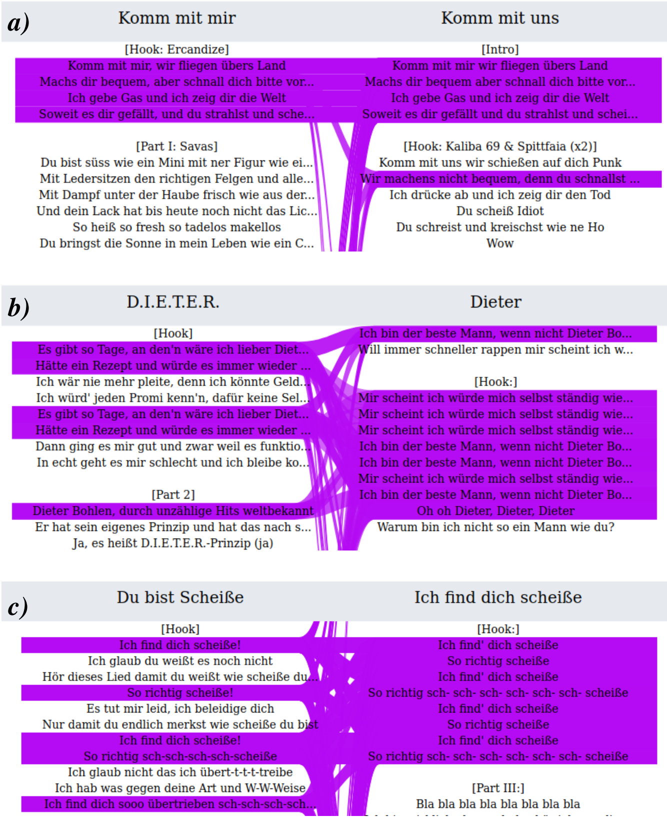

Fig. 4. Excerpt of multiple monolingual alignments on the line-level a) “Kool Savas - Komm mit mir” and “Alligatoah - Komm mit uns”, b) “Fatoni - D.I.E.T.E.R.” and “Samy Deluxe - Dieter” and c) “Sido - Du bist Scheiße” and “Tic Tac Toe - Ich find dich scheiße” .

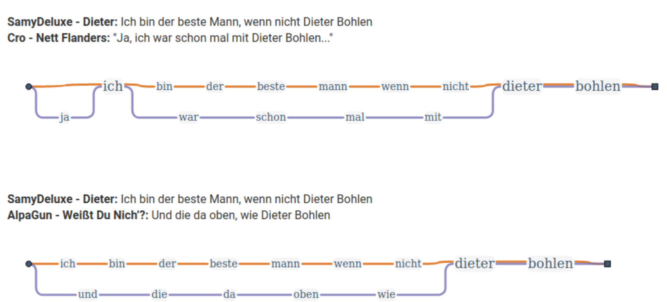

Fig. 5. Two of the nearest neighbors of a line by Samy Deluxe displayed with TraViz.

V. Limitations & Future Works

A. Limitations

A limitation of our approach is the data itself, although Genius always had a strong focus on rap music there are probably always songs or artists that are not included. To increase the knowledge base other information sources could be crawled and linked to the data from Genius. A problem with the alignments is that they are often occurring because of the use of the same proper names like the artist's names or cities, and the usage of the same adlibs. Another problem of the approach is that often, a reference is created through metaphors, rhyme structures or rearrangements of lines, which are hard to detect for automatic methods. To further improve the natural language processing pipeline fine-tuning a model on the urban dictionary [41] could be helpful in order to find more semantic similarities like metaphors. Furthermore, the combination of lyrics and sound features for similarity analysis is of interest. Similar to Yu et al. [23] sound features can be included next to the lyrics to create a multi-modal approach that includes similarities for example in mood, melody, tempo, or rhythm.

B. Multilingual Alignments

We want to extend this approach from monolingual lyrics to multilingual lyrics to detect cases where for example German artists reused passages from American artists. For this, we added the data of around 20 international artists to find multilingual alignments between their lyrics and the lyrics of the German artists as proof of concept. We applied the pre-trained LASER [42] model for 93 different languages to create multilingual sentence embeddings. The LASER encoder maps similar sentences of different languages to similar vectors and can be used without any additional fine-tuning. An alignment, in this case, can be seen as a translation. We found some initial results where the German artists communicated that they reused parts of English songs. Furthermore, the approach is expandable to all music genres and the whole Genius database with over 12 million lyrics. A possible future work would be therefore to use all the data from Genius to detect multilingual references and to compare the similarity of songs based on their lyrics on a large scale through new Distant Reading methods. For example with the visualization of alignments beyond the line-level to inspect multiple texts at the same time or cross-line connections.

C. Future Works

Another interesting approach to the whole corpus would be to display the development of famous quotes like “Each one teach one” over time, cross music genres and beyond music lyrics to other parts of cultural heritage like literature. For this, a citation Timeline would be helpful. Another direction that can be of interest is an analysis of the similarity of the used vocabulary of the artists. This can be supported by stylometry methods, which uses frequencies of uncommon words like the Burrows Delta [43] to find lyrics that are unusual for a given artist and that are more similar to the lyrics of another artist and thus can serve as an indicator for ghostwriting. Such ghostwriters are often not communicated to the audience: “the silent pens might sign confidentiality clauses, appear obliquely in the liner notes, or discuss their participation freely” [44].

VI. Conclusion

We propose a prototype to compute the similarities of rap artists and to find intertextuality between monolingual song lyrics based on word and sentence embeddings. The analysis is supported by visualizations to explore similarities between the lyrics of rap artists. The investigation of the lyrics is further supported by different views showing the metadata from Genius and visualizing similar songs or lyrics through stream graphs to find similar songs and to investigate monolingual alignments in their lyrics. We explained the current limitations of the system, which we noticed through user studies. Furthermore, we outlaid possible directions to focus on, like finding multilingual alignments on a large corpus of song lyrics and cross-modality.

References

[1] G. M. G. Inc., "Genius.com," 2014, https://genius.com/ (Accessed 2020-10-27).

[2] B. N. Kopano, "Rap music as an extension of the black rhetorical tradition:" keepin'it real"," *Western Journal of Black Studies*, vol. 26, no. 4, p. 204, 2002.

[3] S. AB, "Top tracks 2019 deutschland," 2008, https://open.spotify.com/playlist/37i9dQZF1DX4HROODZmf5u (Accessed 2020-10-27.

[4] Y. LLC, "Youtube," 2005, https://www.youtube.com (Accessed 2020-10-27).

[5] S. AB, "Spotify," 2008, https://www.spotify.com/ (Accessed 2020-10-27).

[6] S. Limited, "Soundcloud," 2007, https://soundcloud.com/ (Accessed 2020-10-27).

[7] T. Yousef and S. Janicke, "A survey of text alignment visualization," *IEEE transactions on visualization and computer graphics*, 2020.

[8] R. Khulusi, J. Kusnick, C. Meinecke, C. Gillmann, J. Focht, and S. Jänicke, "A survey on visualizations for musical data," in *Computer Graphics Forum*. Wiley Online Library, 2020.

[9] F. Vavrille, "LivePlasma," 2017, http://www.liveplasma.com/ (Accessed 2020-10-27).

[10] Spotify, "Spotify Artist Explorer," 2018, https://artist-explorer.glitch.me/ (Accessed 2020-10-27).

[11] M. Gibney, "Music-Map," 2011, https://www.music-map.de ((Accessed 2020-10-27).

[12] P. Cano and M. Koppenberger, "The emergence of complex network patterns in music artist networks," in *Proceedings of the 5th international symposium on music information retrieval (ISMIR)*. Citeseer, 2004, pp. 466–469.

[13] M. D. Gleich, L. Zhukov, and K. Lang, "The world of music: Sdp layout of high dimensional data," *Info Vis*, vol. 2005, p. 100, 2005.

[14] M. Daniels, "The largest vocabulary in hip hop," 2014, https://pudding.cool/projects/vocabulary/ (Accessed 2020-10-27).

[15] K. Schramm, "Wer hat den größten?" 2015, https://story.br.de/rapwortschatz/ (Accessed 2020-10-27).

[16] M. D. The Data Face, "The language of hip hop," 2017, https://pudding.cool/2017/09/hip-hop-words/ (Accessed 2020-10-27).

[17] M. Schedl, P. Knees, and G. Widmer, "Discovering and visualizing prototypical artists by web-based co-occurrence analysis." in *ISMIR*, 2005, pp. 21–28.

[18] S. Lu and J. Akred, "History of Rock in 100 Songs," 2018, https://svds.com/rockandroll/#thebeatles (Accessed 2020-10-27).

[19] S. Jänicke, J. Focht, and G. Scheuermann, "Interactive visual profiling of musicians," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 200–209, 2016.

[20] B. Logan, A. Kositsky, and P. Moreno, "Semantic analysis of song lyrics," in *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, vol. 2. IEEE, 2004, pp. 827–830.

[21] S. Baumann and O. Hummel, "Using cultural metadata for artist recommendations," in *Proceedings Third International Conference on WEB Delivering of Music*. IEEE, 2003, pp. 138–141.

[22] R. P. Ribeiro, M. A. Almeida, and C. N. Silla Jr, "The ethnic lyrics fetcher tool," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 27, 2014.

[23] Y. Yu, S. Tang, F. Raposo, and L. Chen, "Deep cross-modal correlation learning for audio and lyrics in music retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, pp. 1–16, 2019.

[24] S. Sasaki, K. Yoshii, T. Nakano, M. Goto, and S. Morishima, "Lyricsradar: A lyrics retrieval system based on latent topics of lyrics." in *Ismir*, 2014, pp. 585–590.

[25] J. Ono, D. Corrêa, M. Ferreira, R. Mello, and L. G. Nonato, "Similarity graph: visual exploration of song collections," in *SIBGRAPI*. IEEE, Institute of Electrical and Electronics Engineers United States, 2015.

[26] R. De Prisco, N. Lettieri, D. Malandrino, D. Pirozzi, G. Zaccagnino, and R. Zaccagnino, "Visualization of music plagiarism: Analysis and evaluation," in *2016 20th International Conference Information Visualisation (IV)*. IEEE, 2016, pp. 177–182.

[27] A. Abdul-Rahman, G. Roe, M. Olsen, C. Gladstone, R. Whaling, N. Cronk, R. Morrissey, and M. Chen, "Constructive visual analytics for text similarity detection," in *Computer Graphics Forum*, vol. 36, no. 1. Wiley Online Library, 2017, pp. 237–248.

[28] S. Jänicke, A. Geßner, M. Büchler, and G. Scheuermann, "Visualizations for Text Re-use," in *Information Visualization Theory and Applications (IVAPP), 2014 International Conference on*. IEEE, 2014, pp. 59–70.

[29] B. Asokarajan, R. Etemadpour, J. Abbas, S. J. Huskey, and C. Weaver, "Textile: A pixel-based focus+ context tool for analyzing variants across multiple text scales." in *EuroVis (Short Papers)*, 2017, pp. 49–53.

[30] C. Di Pietro and R. R. Del Turco, "Between innovation and conservation: The narrow path of user interface design for digital scholarly editions," *Bleier, Klug, Neuber, and Schneider*, pp. 133–63, 2018.

[31] P. Riehmann, M. Potthast, B. Stein, and B. Froehlich, "Visual assessment of alleged plagiarism cases," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 61–70.

[32] S. Jänicke and D. J. Wrisley, "Interactive visual alignment of medieval text versions," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2017, pp. 127–138.

[33] S. Jänicke, A. Geßner, G. Franzini, M. Terras, S. Mahony, and G. Scheuermann, "Traviz: A visualization for variant graphs," *Digital Scholarship in the Humanities*, vol. 30, no. suppl_1, pp. i83–i99, 2015.

[34] P. Riehmann, H. Gruendl, M. Potthast, M. Trenkmann, B. Stein, and B. Froehlich, "Wordgraph: Keyword-in-context visualization for netspeak's wildcard search," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 9, pp. 1411–1423, 2012.

[35] R. H. Dekker and G. Middell, "Computer-supported collation with collatex: managing textual variance in an environment with varying requirements," *Supporting Digital Humanities*, vol. 2, 2011.

[36] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[37] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[38] K. Ethayarajh, "Unsupervised random walk sentence embeddings: A strong but simple baseline," in *Proceedings of The Third Workshop on Representation Learning for NLP*, 2018, pp. 91–100.

[39] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019.

[40] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE symposium on visual languages*. IEEE, 1996, pp. 336–343.

[41] A. Peckham, "Urban dictionary," 1999, https://www.urbandictionary.com/ ((Accessed 2020-10-27).

[42] M. Artetxe and H. Schwenk, "Margin-based parallel corpus mining with multilingual sentence embeddings," *arXiv preprint arXiv:1811.01136*, 2018.

[43] J. Burrows, "'delta': a measure of stylistic difference and a guide to likely authorship," *Literary and linguistic computing*, vol. 17, no. 3, pp. 267–287, 2002.

[44] H. Cameron, "Diddy's little helpers," 2016, https://www.villagevoice.com/2006/11/14/diddys-little-helpers/ (Accessed 2020-10-27).