Visualizing Similarities between American Rap-Artists

C. Meinecke,[1] J. Schebera,[1] J. Eschrich[1] and D. Wiegreffe[1]

[1]Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Leipzig, Germany

Abstract

Rap music is one of the biggest music genres in the world today. Since the early days of rap music, references not only to pop culture but also to other rap artists have been an integral part of the lyrics' artistry. In addition, rap musicians reference each other by adopting fragments of lyrics, for example, to give credit. This kind of text reuse can be used to create connections between individual artists. Due to the large amount of lyrics, only automated detection methods can efficiently detect similarities between the songs and the artists. Here, we present a visualization system for analyzing rap music lyrics. We also trained a network tailored specifically for rap lyrics to compute similarities in lyrics. Here a video of the prototype can be seen.

CCS Concepts

• ***Human-centered computing*** → ***Visualization application domains; Visualization systems and tools;*** • ***Applied computing*** → ***Sound and music computing; Document management and text processing;***

1. Introduction

Rap music started as a way for marginalized groups to express their social and economical struggles rhythmically and poetically. In the early years after its inception, the genre stayed mostly within the borders of its corresponding subculture. But in the 1980s, with the emergence of "gangsta rap" through groups like N.W.A and artists like Snoop Dogg or Dr. Dre, rap music made its breakthrough into the mainstream [AS05, Lig99]. Today, it is one of the biggest music genres with its influence spanning across the globe [mos] and Websites like Spotify [AB08] provide access to millions of songs on demand while Genius.com [Inc14] offers annotated song lyrics. Since rap music's early days, references to pop culture but also to other rap artists have been an integral part of its lyrical craftsmanship. Rappers may share personal connections through their backgrounds like the city or neighborhood they grew up in. Because of these relations they often reference similar themes, places, or culturally specific phrases. References can also be formulated in a negative way. Rivalries spanning the whole genre like the East Coast vs. West Coast clash in the 1990s often result in so-called "diss tracks". In these, the musicians mock each other, often re-using or referencing their adversary's lyrics to use against them. More positively, artists sometimes re-use other musicians' phrases to pay homage to them and their lyrical craftsmanship. Because of the sheer amount of lyrical content, automated means of detecting similar artists based on their lyrical content can help fans of the genre to deepen their knowledge. The result of this work is a visualization system that enables the user to explore similarities between artists, detect allusions between songs, and discover new artists or songs. The lyrics are embedded using RoBERTa [LOG*19] in order to compute similarities between them. The system can be also applied to lyrics from other genres, but this work focuses on rap as a case study because of the already mentioned multitude types of references integrated into the genre.

2. Related Works

Previous works have utilized visualization to compare artists based on user statistics and reviews, and also to show plagiarism [KKM*20]. Similar to us, Meinecke et al. [MHJ22] use lyric data to generate similarities between artists. However, our methods are fine-tuned on the semantic textual similarity task and on a corpus of rap lyrics to include domain knowledge.

3. Tasks

The application is designed for users of the general public interested in rap music with the aim to offer a tool that supports an exploratory analysis of selected American rap musicians and their lyrics. Therefore, tasks were developed by the authors for the design of this application, which follow a finer and finer order from artists to songs to lines. Thereby, each task can serve as an entry point for one another, but can also be treated separately in the tool. On an artists level, a user could want to discover artists similar to those they are already familiar with (Task 1.1. **Find similar artists**), to explore the artists' background (Task 1.2. **Explore an artists**), or even explore groups of artists to find artists directly referencing each other (Task 1.3. **Compare different artists**). On a song level, users may also be interested in finding lyrically similar songs to a song of interest (Task 2.1. **Find similar songs**), to explore by searching for other songs that reference specific lines (Task 2.2. **Explore a song**), or when a user found a song of interest they

DOI: 10.2312/evp.20221129

could be interested in comparing the song to another side-by-side (Task 2.3. **Compare different songs**). Finally, a user could be interested to find lines that are similar to a line of interest and finding all occurrences across the corpus (Task 3. **Find similar lines**).

4. Data & Processing

We collected data about 219 American hip hop artists based on popularity and personal interest with a maximum of 200 songs per artist from Genius.com [Inc14]. The data contain information about the artists, like their names, a short description, and the artists' songs including their lyrics. Since relationships and similarities are our main focus, we used a Neo4j graph-database to store artists, songs and lines as nodes. The structure of of this database is illustrated in Figure 1. We focus on similarities between individual lines to establish connections between songs and artists. For this, each line in a song is represented by its own node in the database. We connect the line-nodes with relationships based on the results of the similarity search. Each song is represented by a node as well, containing information about the title, release date, associated album, featured artists, etc. Line-nodes are connected to their respective song-nodes via a "part-of" relationship. Finally, the same is done for the artists, as those too are represented by their own nodes containing their name, description, alternate names, etc.

In order to find lines that are semantically similar, we used RoBERTa [LOG*19]. The model takes a sentence as input and produces an embedding vector representing the semantic meaning. We utilized two versions, one ready-to-use version specifically fine-tuned on the task of semantic textual similarity called 'stsb-roberta-base', and the same network additionally fine-tuned on our collected corpus of rap lyrics which we gave the name 'rapBERTa'. The reason for this additional training is the heavy usage of slang, neologisms and pop culture references in hip hop. The hypothesis was that in learning rap-specific language, rapBERTa may also perform better in finding meaningful semantic similarities in a corpus of rap lyrics. The vector of each line is indexed using faiss [JDJ19] for efficient similarity search. The index was used to find the 15 nearest neighbors for each line i.e. the most similar lines within the corpus based on cosine similarity. The resulting neighbor relations between song lines are then added to the Neo4j graph-database with their corresponding similarity value as "neighbor of" relationships. It should be mentioned that the similarities between artists and songs are also calculated based on the corresponding average line similarities.

5. Visual Interface

The user can explore the acquired data through a web application that provides several visualizations to aid in discovering patterns and relationships. The *artist graph* shown in Figure 2 represents the core of the application. This force-directed graph layout provides the user an overview of all the artists in the database and their similarities (Task 1.1.). Each artist is represented by a circle containing an image of the artist. An edge between two artists indicates that they are the most similar based on their lyrics. This leads to the formation of subgraphs consisting of lyrically related artists. The connections between the artists within the subgraphs help the user quickly identify groups of similar artists. With this baseline of information, the user can then explore the lyrical connections of artists within these groups (Task 1. & 2.). From the artist graph, the user can select an artist. This opens a popup (*artist view*) containing information about the artist and a list of their songs, which supports Task 1.2. ("Explore an artist"). Selecting an additional artist will open a second artist view on the right side, which can be seen in Figure 3. At the top of the artist view, the user can find a short text about the artist.

Opening two artist views offer the first method of direct comparison Task 1.3. ("Compare different artists"). An additional popup in the middle between the two artist views, shows pairs of the artists' most similar songs. This also allows to support Task 2.1. ("Find similar songs"). Selecting one of these pairs will open a *song view*, in place of their corresponding artist view. Whenever there are two song views open at the same time, all their similar lines are shown by colored Bézier curves (Task 3.). This visualization (Figure 4) can be thought of as a graph, where the song lines are vertices with edges connecting them to similar song lines. Each connected component in this graph has its own color, so the user can easily differentiate between groups. The songs remain individually scrollable, so different parts of both songs can be compared and explored (Task 2.2. & 2.3.). If the user wants to find references to a song, opening only one song view shows a list of similar songs in the middle of the screen (Task 2.2.). To get even more specific, each line in a song view is clickable. Selecting one line opens a list of all similar lines from other songs on the opposite side of the screen (Task 3. "Find similar lines"). This enables the user to explore the usage of certain phrases between different artists and possibly trace who is referencing who. Additionally, a Text Variant Graph (Figure 5) aids in the comparison of similar lines. Each word, or group of words that the language model deemed similar, is represented by a box. Colored arrows connect the boxes to form the sequences of words as they appear in the song lines. Each path of one color represents one song line. The song line selected as consensus is displayed in the center as a sequence of nodes aligned horizontally on a line (red edges). Thus, the sequences diverge where the choice of words differs between the lines, and converge where the chosen variant words are the same.

6. Future Works

Training on a larger corpus of lyrics could improve the results. Moreover, a manually assembled data-set of similar and dissimilar lines could be used to fine-tune the model. This could be supported by a Visual-Interactive Labeling approach [RAZ*18] for example in a crowd-sourced environment. Other possibilities are to include multi-modal information like music samples, using multilingual similarities [AS19] or the detection of similarities across music genres.

7. Conclusions

The proposed prototype offers visual tools for users interested in rap music to discover similarities between rap artists and their lyrics and to explore a corpus of rap lyrics through visualizations. Two sentence transformers produce sentence embeddings for each individual line to automatically detect semantically related lines.

References

[AB08] AB S.: Spotify, 2008. https://www.spotify.com/ (Accessed 2021-10-27). 1

[AS05] ALRIDGE D. P., STEWART J. B.: Introduction: Hip hop in history: Past, present, and future. *The Journal of African American History 90*, 3 (2005), 190–195. URL: http://www.jstor.org/stable/20063997. 1

[AS19] ARTETXE M., SCHWENK H.: Margin-based parallel corpus mining with multilingual sentence embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), 3197–3203. 2

[Inc14] INC. G. M. G.: Genius.com, 2014. https://genius.com/ (Accessed 2021-10-27). 1, 2

[JDJ19] JOHNSON J., DOUZE M., JÉGOU H.: Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* (2019). 2

[KKM*20] KHULUSI R., KUSNICK J., MEINECKE C., GILLMANN C., FOCHT J., JÄNICKE S.: A survey on visualizations for musical data. In *Computer Graphics Forum* (2020), Wiley Online Library. 1

[Lig99] LIGHT A.: *The Vibe History of Hip Hop*. Three Rivers Press, 1999. URL: https://openlibrary.org/books/OL42726M/The_Vibe_history_of_hip_hop. 1

[LOG*19] LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L., STOYANOV V.: Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019). 1, 2

[MHJ22] MEINECKE C., HAKIMI A. D., JÄNICKE S.: Explorative visual analysis of rap music. *Information 13*, 1 (2022), 10. 1

[mos] Hip-Hop Becomes Most Popular Genre In Music For First Time In U.S. History – VIBE.com. URL: https://www.vibe.com/music/music-news/hip-hop-popular-genre-nielsen-music-526795/. 1

[RAZ*18] RITTER C., ALTENHOFEN C., ZEPPELZAUER M., KUIJPER A., SCHRECK T., BERNARD J.: Personalized visual-interactive music classification. In *EuroVA@ EuroVis* (2018), pp. 31–35. 2

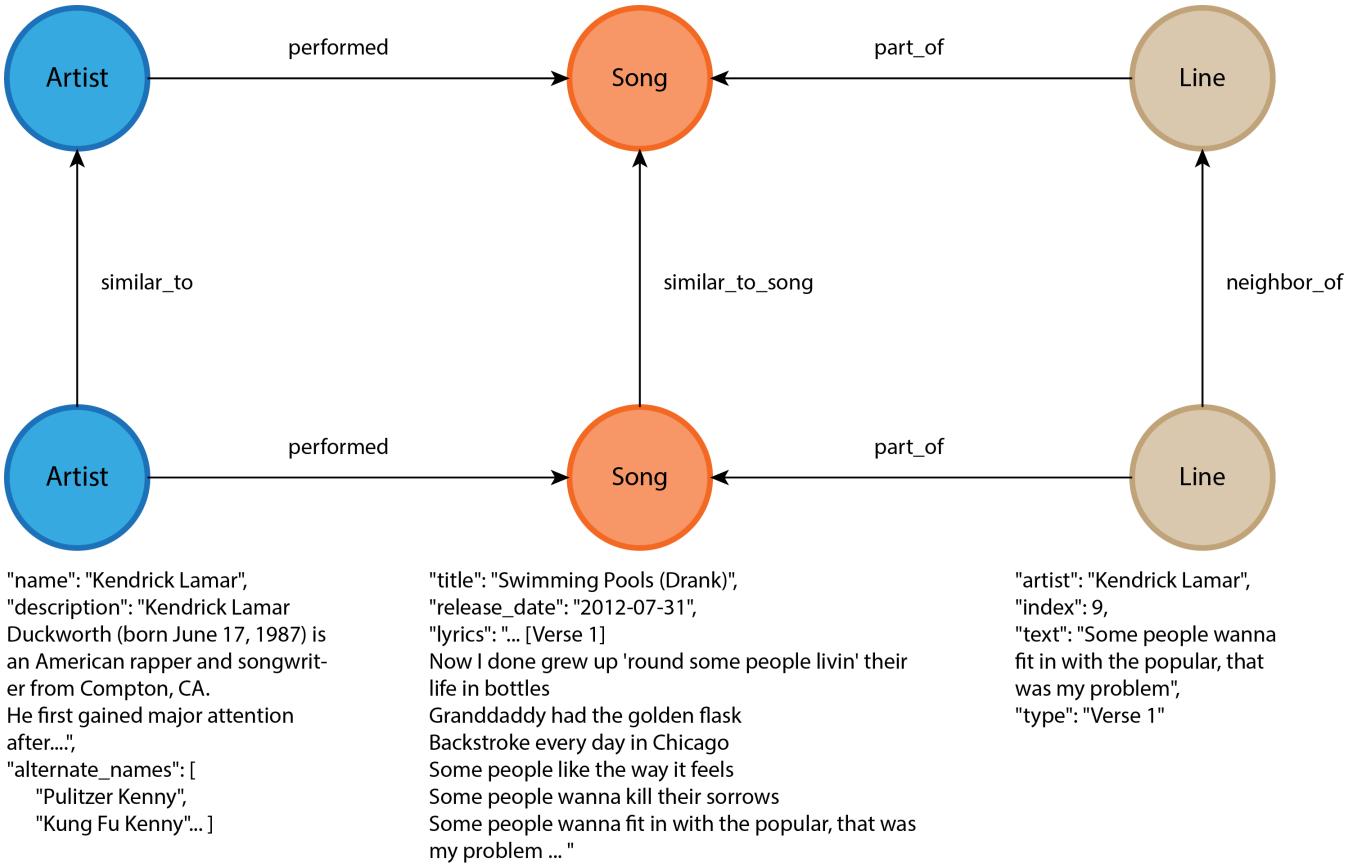

Figure 1: *The structure of our Neo4j graph-database is illustrated here with a small example with artists, songs and lines as nodes. Edges are established between nodes to create real links like artists perform songs and lines are part of a song. Also, based on the line similarities, similar lines, songs and artists are connected by edges.*

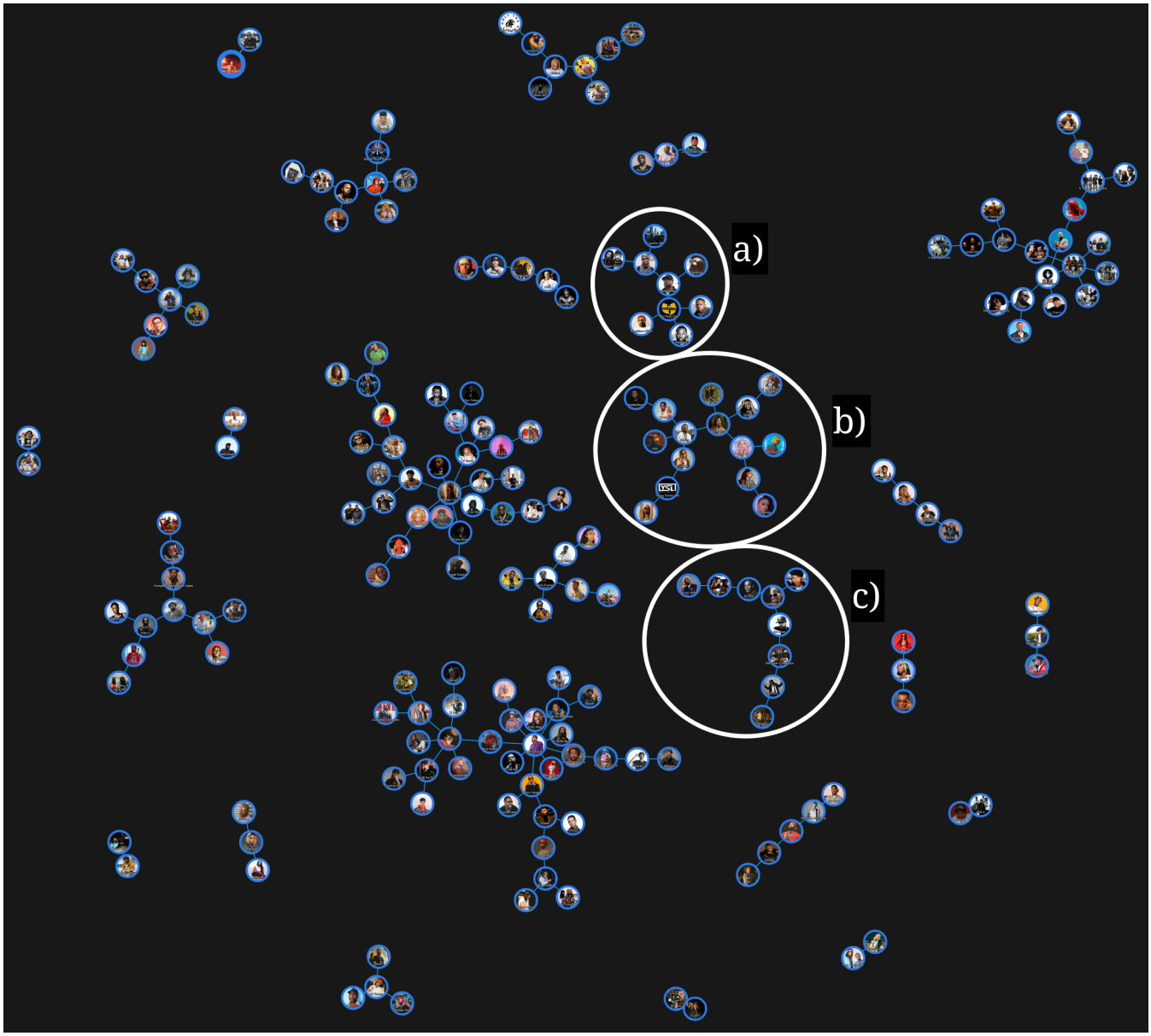

Figure 2: *The artist graph, artists that are similar based on their lyrics are connected. Different kinds of clusters can be observed.* ***a)*** *Subgraph containing the artists Raekwon, Ghostface Killah, Method Man, Redman, and GZA, all part of the Wu-Tang Clan. The additional artists featured in the subgraph, also emerged in the same time period around 1990.* ***b)*** *Subgraph with a cluster containing Atlanta based rappers Offset, Quavo and Take-off. They form the rap trio known as 'The Migos'. We can also see a close connection between Offset and Cardi B, who are married in real life. Three of the other rappers in this subgraph are also based - or at least born in - Atlanta.* ***c)*** *Shows N.W.A members Dr. Dre and Ice Cube together with several artists connected to them. Including Westside Connection a group where Ice Cube was a member of, Snoop Dog and Warren G two artists that collaborated with Dr. Dre, and House of Pain were one of the members was part of Ice Cubes collective Rhyme Syndicate.*

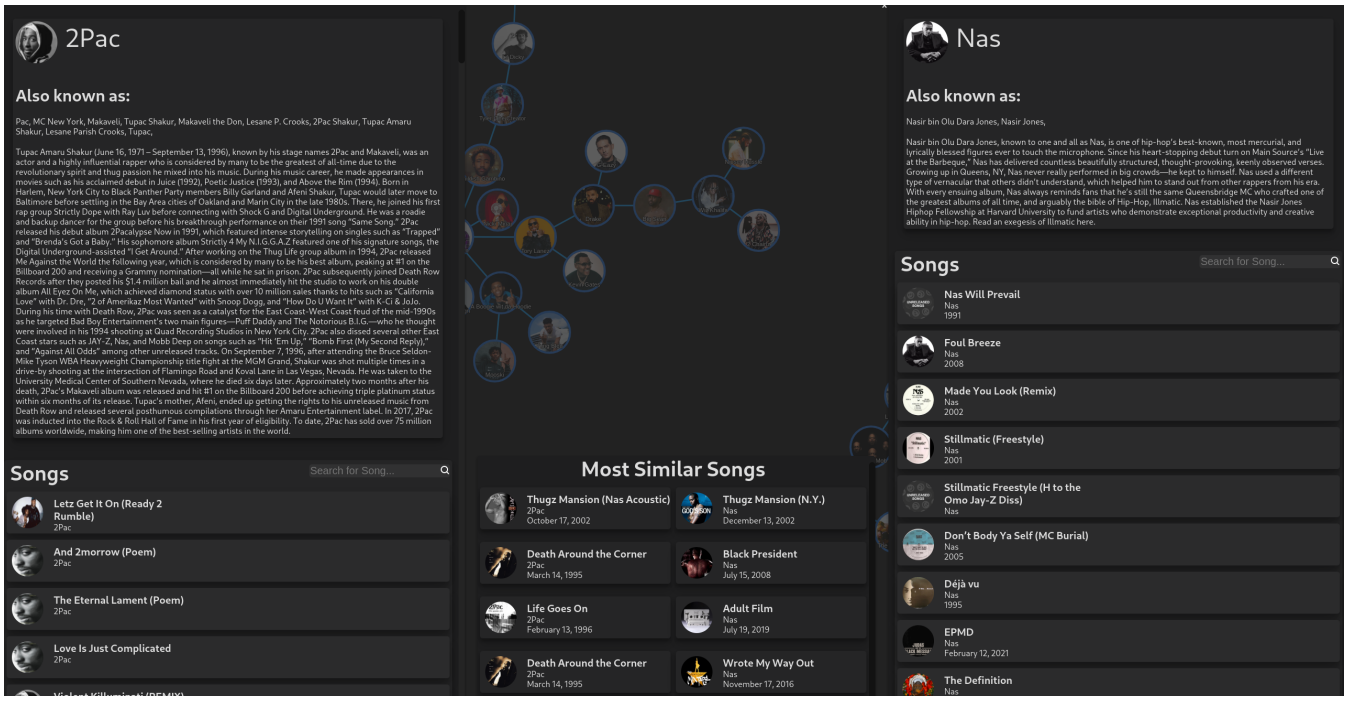

Figure 3: *The artist view of 2Pac and Nas shows biographical information, a list of the songs and the most similar songs of both artists.*

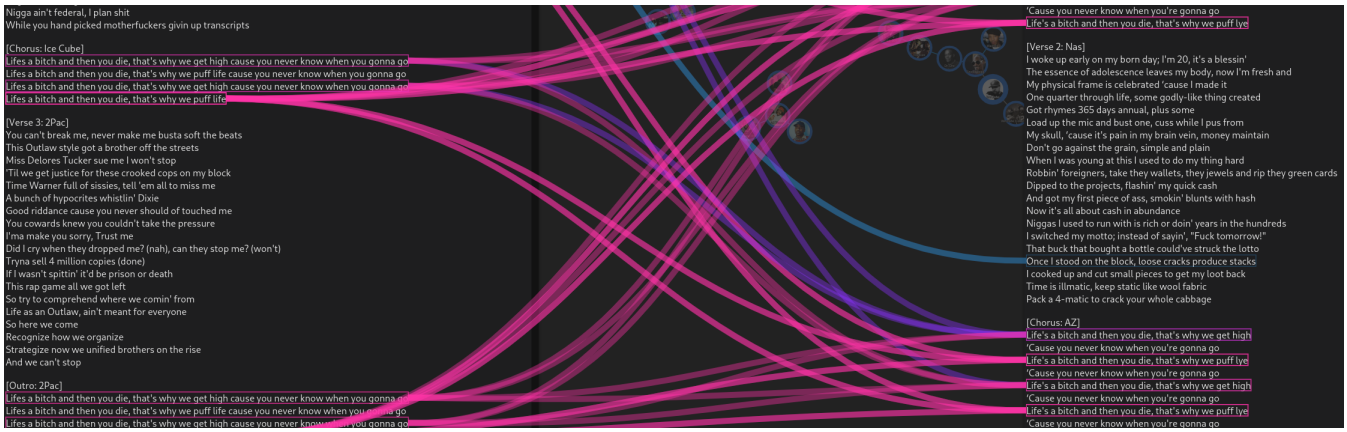

Figure 4: *Side by Side View of the songs "Fear Nothing" by 2Pac and Ice Cube and "Life's a Bitch" by Nas and AZ. The former song reused the chorus of the later song. Each group of lines that are similar to each other is assigned a unique color so that the user can easily distinguish them.*

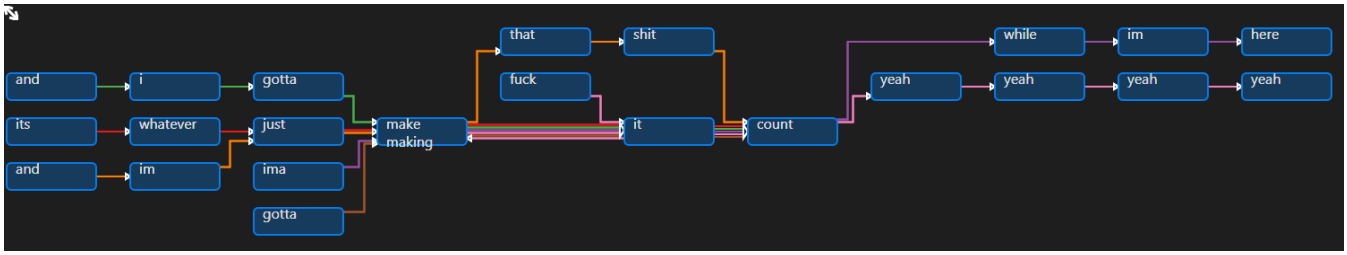

Figure 5: *The Text Variant Graph, the different lines are color coded and shared words are merged into one node.*

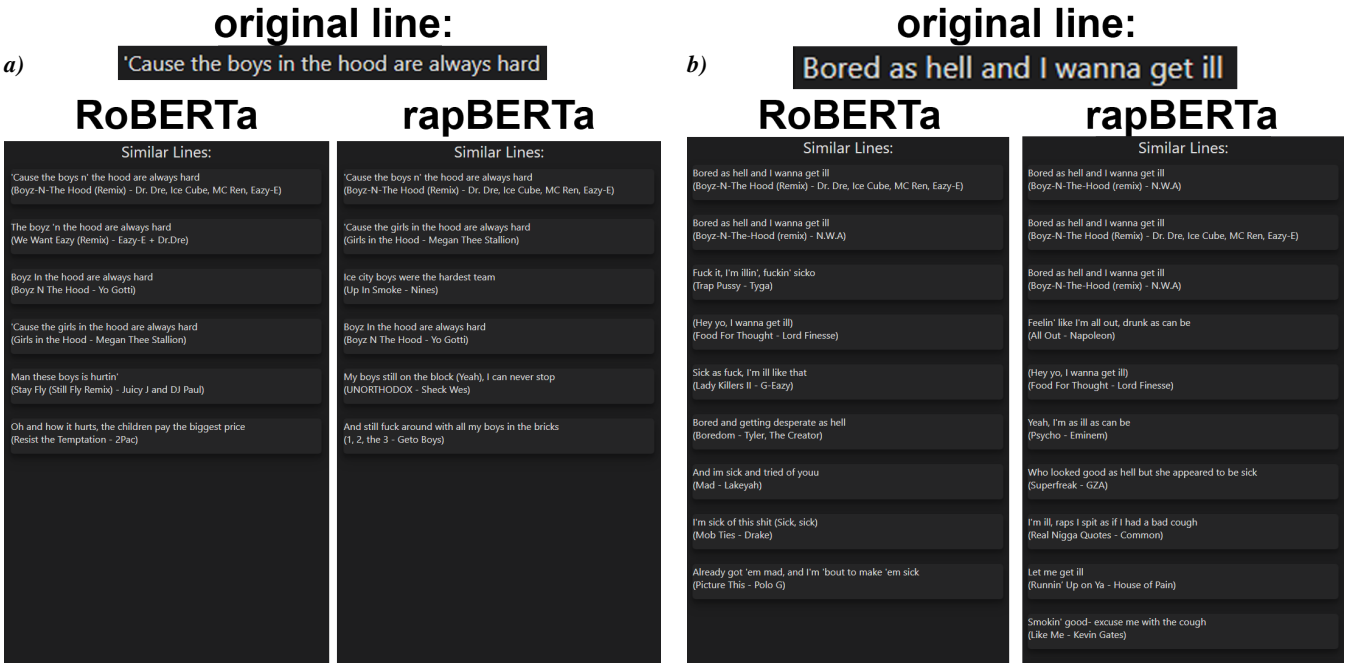

Figure 6: *The most similar lines based on the two different models for the line* ***a)*** *"Cause the boys in the hood are always hard" and* ***b)*** *"Bored as hell and I wanna get ill".*

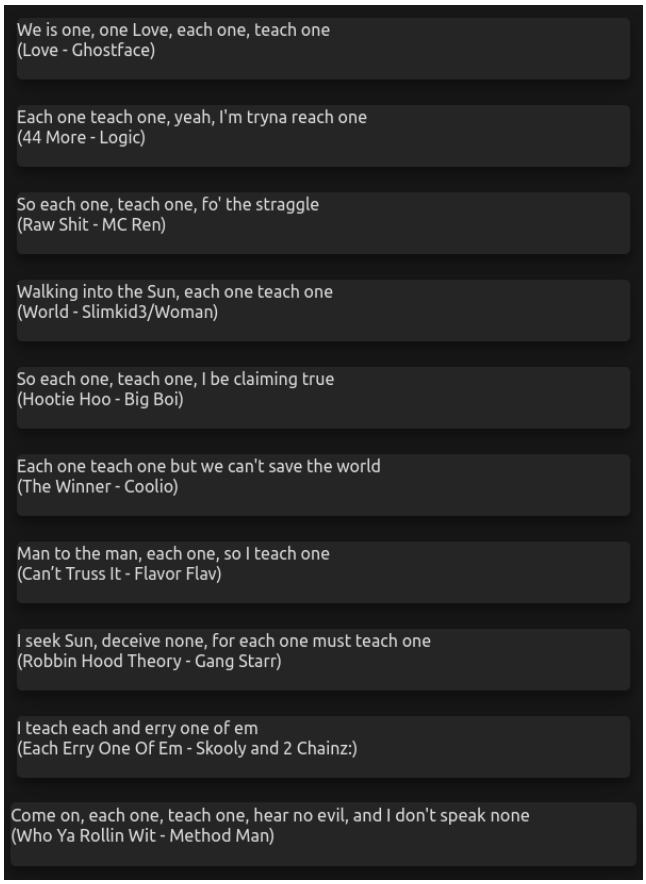

Figure 7: *Top search results for the African-American proverb "Each one teach one".*