Classification and Visualization of Lyric Collections Using Guided LDA

1st Yuki Nakai
Graduate school of humanities and sciences
Ochanomizu University
Tokyo, Japan
g1820524@is.ocha.ac.jp

2nd Takayuki Itoh
Graduate school of humanities and sciences
Ochanomizu University
Tokyo, Japan
itot@is.ocha.ac.jp

***Abstract*—Lyrics are one of the most important components of music and have a great impact on our daily listening to music. Music search based on the meaning of lyrics is therefore useful. However, the impression of lyrics is subjective, and therefore, it is often difficult to deal with music so that every listener feels satisfied while focusing only on the lyrics. Based on this discussion, we aimed to develop an exploratory user interface that enables users to subjectively explore the music based on their own impressions of lyrics. This paper proposes a method to interactively visualize the distribution of lyrics by applying guided LDA (Latent Dirichlet Allocation).**

***Index Terms*—visualization, lyric, LDA, guided LDA**

I. Introduction

Lyrics have a great impact on the appreciation of songs such as pop music since they are one of the most important components of music. Lyrics-based music search and classification is therefore useful. However, the impression of lyrics is subjective and may be influenced by musical elements other than lyrics, so the criteria for searching for lyrics required by users may vary from person to person. To address this issue, we are working on a research project to support exploratory lyric searches by visualizing the distribution of lyrics. Here, it is often difficult to appropriately calculate the distribution of the lyrics because lyrics have a higher degree of lexical freedom than articles and academic papers. In this study, we propose a method to visualize the distribution of lyrics calculated applying guided LDA (Latent Dirichlet Allocation) that interactively consumes guided words. This method facilitates the iterative visualization of lyric classification results based on the users' viewpoints. We aim to develop a user interface for searching for songs by focusing only on lyrics without taking other musical elements into account. Users can observe the differences in individuality and tendency of songs and artists, and the diversity of lyrics, by using the visualization results.

We suppose the target users of this study as music listeners, music industry workers, and songwriters. Music listeners can use the visualization results to search for songs whose lyrics are similar to their favorite songs as a starting point. Music industry workers can also use it to discover the individuality of the artists along with the lyrics, or to get an overview of lyrics trends. Furthermore, lyricists can use this method as a reference in their search for lyric writing methods that are unique to each artist. The visualization with dimensionality reduction and scatterplots adopted in this study is suitable for an overview of the entire distribution of the lyric data. This feature is mainly suitable, for example, for those who work in the music industry and want to get an overview of the overall trend of hit songs. Meanwhile, ordinary music listeners often want to search for songs locally, for example, to search for songs similar to those of a particular artist or to search for songs with lyrics that contain specific content. In such cases, users of this visualization can discover by zooming in on a specific part that the user is interested in, starting from an overview of the distribution of lyrics.

Since the motivation of this study is to archive the exploratory search of Japanese hit songs (so-called J-POP), this paper presents examples of visualizations with a lyrics dataset of J-POP songs.

The reminder of this paper is as follows. Section 2 introduces related studies, Section 3 describes the proposed method, Section 4 shows an example of the implementation of the method, and Section 5 summarizes the study and discusses future issues.

II. Related work

This section introduces related studies on analysis and visualization of the lyrics of J-POP songs since our motivation is the visualization of the distribution of J-POP songs.

Kobayashi et al. [1] compared the usage rates of lexical indicators such as part of speech and word type for words used in lyrics, and found that time-series changes in language use were analyzed.

Ohde et al. [2] explored the chronological trends and culture of lyrics by dividing words into groups, registering codes for each group, and analyzing the words that frequently appear in lyrics and the corresponding codes. Sadamura et al. [3] used quantitative text analysis of lyrics of songs written by Matsumoto Takashi, and compared the frequency of use by different singers.

Kawamura [4] proposed a method for recommending songs based on search words and their associative words by automatically extracting associative words of search words and calculating features of lyrics using the TF-IDF method. Hosoya et

al. [5] conducted an exploratory analysis of lyrics of multiple female singer-songwriters using a random forest.

Hossain et al. [6] proposed a method to analyze the lyrics of a large number of songs using LDA and to recommend song titles based on the lyrics. Sasaki et al. [7] proposed a user interface for lyrics search called "LyricsRadar," which uses LDA to determine the latent meaning of lyrics and allows users to interactively search for their preferred lyrics from a large number of registered songs.

The proposed technique differs from the above studies since we apply guided LDA to calculate the distribution of lyrics. It also differs from the studies since our technique can provide various explanations for the distribution and similarity of songs by displaying multiple scatterplots for each set of guide words.

III. Proposed visualization

A. Collection of music data

Our dataset contains the title, artist, lyricist, date, and lyrics for a song. We collected the dataset on the top 10 highest-selling CD singles from 1988 to 2007 and the top 10 songs on the Billboard Japan annual chart from 2008 to 2020. Here, we excluded songs that consist only of English lyrics from the dataset. For songs that appeared in multiple years, we recorded only the most recent data to avoid duplication of the song data. We recorded a total of 332 songs as a result.

Our implementation uses MeCab for the morphological analysis of lyrics. We extract only nouns, verbs, adjectives, adverbs, inspirations, and coordinating verbs, and count their prototypes as one word. Here, we determined "non-important words" that are words appearing frequently in an extremely large number of songs. This study supposes to set such words as "stop words" and exclude them from the analysis.

B. Analyzing lyrics

1) Analyzing lyrics using LDA: LDA [8] is a model that assumes that a set of documents consists of multiple topics. Figure1 shows the graphical model of LDA. The manifest variables are represented by shaded vertices, latent variables and unknown parameters by other vertices, and their dependencies by directed edges. The rectangles indicate that the creation of variables within the rectangle is repeated as many times as the numbers marked in the corners. This method assumes that each lyric is an independent document and uses D independent lyrics $X = \{X_1, \ldots, X_n\}$. The lyric X_d consists of N_d words $X_d = \{w_{d,1}, \ldots, w_{d,N_d}\}$, and this method assumes that each word in each lyric has latent topics. K is the number of topics, θ is the topic multinomial distribution parameter for each document, ϕ is the word multinomial distribution parameter for each topic, and α and β are the Dirichlet hyperparameters for θ and ϕ respectively.

The process of document generation from a graphical model is shown below.

1) Select $\theta_d \sim Dir(\alpha)$ for each document X_d
2) Select $\phi_k \sim Dir(\beta)$ for each topic k
3) For each of the N_d words $w_{d,i}$ in document X_d
   - Select topic $z_{d,i} \sim Mult(\theta_d)$
   - Select word $w_{d,i} \sim Mult(\phi_{z_{d,i}})$

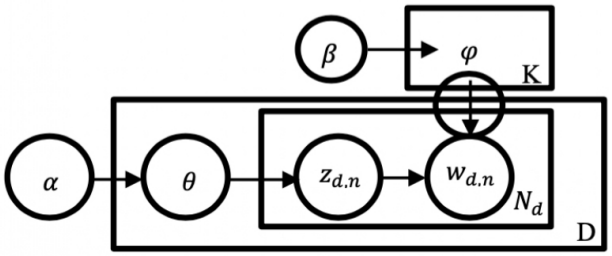

Fig. 1. Graphical model of LDA.

This method uses Gibbs sampling [11] to estimate the topic model. Assuming that everything is fixed except for the topic assigned to the i-th word in X_d document , the probability that the topic is k is given by

$$P\left(z_{d,i} = k \mid w_{d,i} = w, z\_i, w\_i, \alpha, \beta\right)$$

$$\propto \frac{C^{KD}_{kd,-i} + \alpha}{\sum_{k'} C^{KD}_{k'd,-i} + K\alpha} \cdot \frac{C^{WK}_{wk,-i} + \beta}{\sum_{w'} C^{WK}_{w'k,-i} + K\beta}$$

Here, w_i is the set w_{-i} excluded from w and z_{-i} is the set z_i excluded from z. $C^{KD}_{kd,-i}$ is the number of times topic k is assigned to document X_d and $C^{WK}_{w'k,-i}$ is the number of times topic k is assigned to word w. The i-th word is excluded in both cases. In this method, the expected value of the topic multinomial distribution parameter θ which is the topic mixture ratio of lyrics is obtained for each lyric, and by applying dimensionality reduction as described below in Section 3.3, the distribution of songs is visualized on a two-dimensional plane.

2) Analyzing lyrics using guided LDA: Guided LDA [9] is an extended LDA that assigns important words as reserved words (guide words) to each topic in advance. This makes it more likely that the words that are classified as representative words for each topic are words that co-occur with the guide word. Therefore, users can introduce clear perspectives into the topic classification results. Here, we obtain the probability values of the topics for each lyric as in LDA. This paper presents the selections of topic themes and guide words according to the following two types.

- season-based topic themes
- event-topic themes

Table I shows the topic themes and topic words employed in this study.

We set the following words shown in TableII that frequently appeared as non-important words in the topics while analyzing all lyrics using LDA as stop words.

For lyrics with mixed Japanese and English lyrics, only the Japanese lyrics were included in the LDA analysis.

In our experiments, we also applied unguided normal LDA in addition to the guided LDA for the comparison of visualization results.

TABLE I
TOPIC THEMES AND GUIDE WORDS IN GUIDED LDA

(1) season-based topic themes

topic theme	guide words
spring	haru (spring), sakura (cherry blossom), sotsugyo (graduation)
summer	natsu (summer), matsuri (festival), hanabi (fireworks), himawari (sunflower)
autumn	aki (autumn)
winter	fuyu (winter), yuki (snow)

(2) event-based topic themes

topic theme	guide words
graduation	sotsugyo (graduation), tomo (friend), wakare (farewell),sayonara (good bye) sakura (cherry blossom), seifuku (school uniform), kadode (departure)
summer festival	natsu (summer), matsuri (festival), hanabi (fireworks), kori-gashi (sherbert)

TABLE II
STOP WORDS

suru (do), aru (be), iru (be), naru (be), nai (not), kono (this), sono (its), ano (that), sore (it), boku (I), watashi (I), ore (I), anata (you), kimi (you), bokura (we)

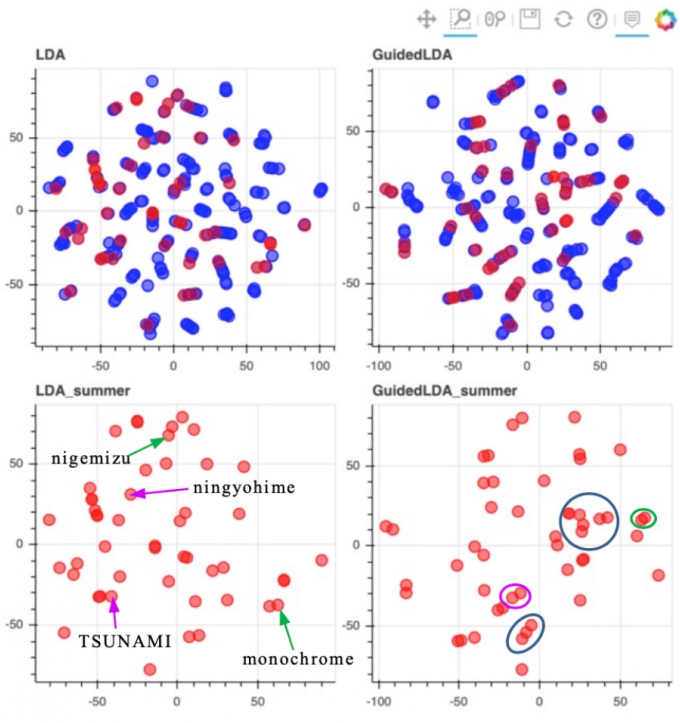

Fig. 2. (upper left) Application of LDA. (upper right) Application of guided LDA. (lower left) Only lyrics including ”natsu” (summer) (LDA). (lower right) Only lyrics including ”natsu” (summer) (guided LDA).

C. Visualization

We visualize the results of the analysis of the lyrics using guided LDA as a scatterplot with dimensionality reduction to two dimensions using t-SNE [10]. Here, a single point in the scatterplot represents a single lyric.

IV. EXAMPLES

This section shows the visualization examples below. We fixed the number of topics as $K = 4$ in the experiment.

A. Season-based topic themes

This paper introduces examples that display the lyrics containing the word ”natsu” (summer) as red dots and the other lyrics as blue dots in the form of a scatterplot as shown in Figure 2. Compared to the case with the LDA, the guided LDA resulted in fewer independent points and the lyrics were placed along with the similarity of the lyrics. We can see that all the lyrics are related to summer and have a romantic theme while looking at the scatterplot with three or more points (see the blue circle in Figure 2). Next, we focused on the two points that are farther apart when LDA is applied, of the two points that overlap in the scatterplot when guided LDA is applied. The two points included in the green circle in Figure2 are ”nigemizu” (mirage) by Nogizaka46 and ”monochrome” by Ayumi Hamasaki. These two songs have a common theme of lost love and the use of the words ”yume” (dream) and ”maboroshi” (illusion) to describe summer love. The two dots in the purple circle in Figure 2 are ”ninghohime” (little mermaid) by Miho Nakayama and ”TSUNAMI” by Keisuke Kuwata. The lyrics of these two songs are related to past summer love, and the word ”ame” (rain) is frequently used in the lyrics of both songs. Thus, the guided LDA is better in terms of the placement of songs that are not only related to summer but also have commonalities in the description of the scene.

This section also shows the visualization results when we set ”natsu” (summer) and ”hanabi” (fireworks) as the guide words for one topic, and no guide words for the other topics, as shown in Figure 3. This figure indicates the lyrics containing the word ”natsu” (summer) by red dots, while indicating specific two songs, ”hanabi” (fireworks) by Mr. Children and ”sayonara-zinrui” (goodbye humanity) by Tama, by green dots, and the other lyrics by blue dots. These two songs include the word ”hanabi” (fireworks) in the lyrics, but the lyrics as a whole are not related to summer. The ”hanabi” (fireworks) is an independent point in both the scatterplots applying LDA and the guided LDA. However, the distance to the nearby point is farther in the case with the guided LDA than in the case with the LDA. For ”sayonara-zinrui” (goodbye humanity), the points are located in the area where summer-related songs such as ”rokorosyon” (loco-lotion) by ORANGE RANGE are concentrated in the scatterplot while using LDA. On the other hand, the point corresponding to ”sayonara-zinrui” (goodbye humanity) is independent on the scatterplot while using the guided LDA. In other words, the guided LDA provides a better separation of the placement of songs related only to ”summer” and songs related only to ”fireworks”.

B. Event-based topic themes

The scatterplot shown in Figure 4 indicates the lyrics that contain the word ”sotsugyo” (graduation) and are related to

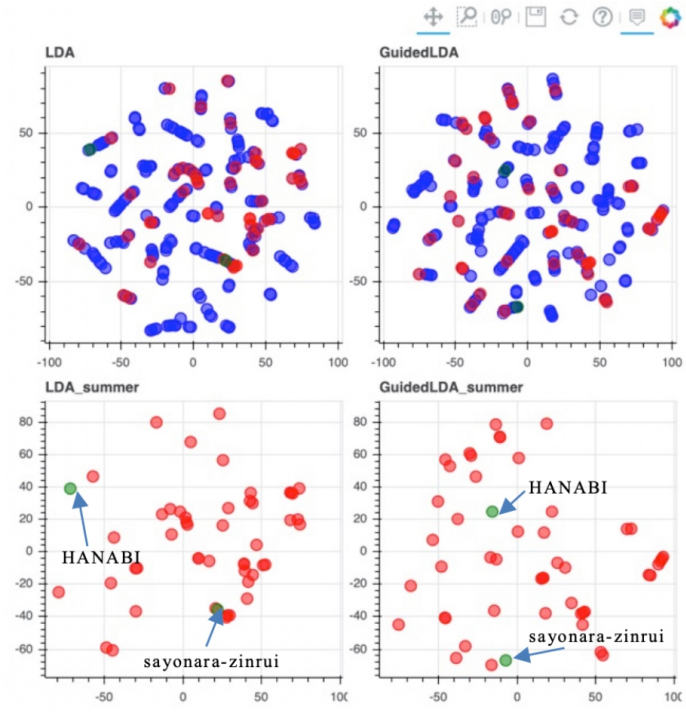

Fig. 3. (upper left) Application of LDA. (upper right) Application of guided LDA. (lower left) Lyrics including ”natsu” (summer) and only ”HANABI” (fireworks) and ”sayonara-zinurui” (goodbye humanity) (LDA). (lower right) Only the lyrics including ”natsu” (summer), ”HANABI” (fireworks) and ”sayonara-zinrui” (goodbye humanity) (guided LDA).

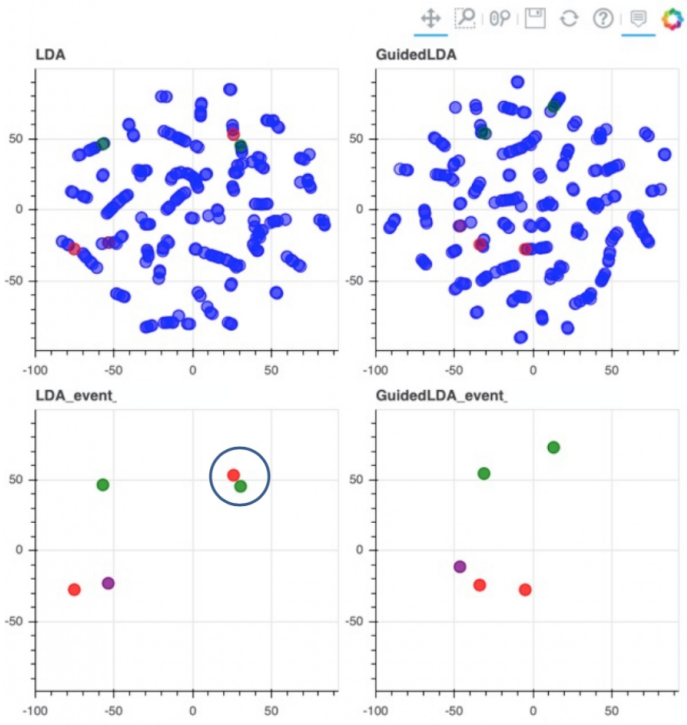

Fig. 4. (upper left) Application of LDA. (upper right) Application of guided LDA. (lower left) Only lyrics including ”sotsugyo” (graduation) and lyrics with content related to graduation (LDA). (lower right) Only lyrics including ”sotsugyo” (graduation) and lyrics with content related to graduation (guided LDA).

graduation by red dots, the lyrics that do not contain the word ”sotsugyo” (graduation) but are related to graduation by purple dots, the lyrics that contain the word ”sotsugyo” (graduation) but are not related to graduation by green dots, and the other lyrics by blue dots. In the scatterplot using LDA, the distance between lyrics that contain the word ”sotsugyo” (graduation) and are related to graduation (red dots) and lyrics that include the word ”sotsugyo” (graduation) but are not related to graduation (green dots) are closer. On the other hand, in the scatterplot using guided LDA, lyrics that contain the word ”sotsugyo” (graduation) and are related to graduation are farther apart from lyrics that contain the word ”sotsugyo” (graduation) but are not related to graduation. In addition, the lyrics related to graduation (red dots and purple dots) are placed close to each other regardless of the presence or absence of the word ”sotsugyo” (graduation). Due to the above, guided LDA is better at separating lyrics related to graduation and lyrics not related to graduation.

The scatterplot shown in Figure5 indicates the lyrics that contain the word ”matsuri” (festival) and are related to summer festival by red dots, the lyrics that contain the word ”matsuri” (festival) but are not related to graduation by green dots, and the other lyrics by blue dots. In both cases of using LDA and guided LDA, lyrics related to summer festivals and lyrics not related to summer festivals are placed far apart on the scatter plots. Meanwhile, we can see that the distance between the points using guided LDA is further than that using LDA.

V. Conclusion

We proposed a method for quantifying lyric topics using guided LDA and displaying them as scatterplots by applying dimensionality reduction as a method for visualizing the distribution of lyrics in this paper. Individual viewpoints of users are introduced into the visualization results of the distribution of lyrics by selecting guide words using this method. This facilitates the interpretation of lyrics’ similarities and differences from the user’s personal perspective. In addition, users can create multiple visualization results by changing the selection of guide words selected, making it easy to compare the distribution of lyrics from various viewpoints. Compared to the non-guided ordinary LDA, guided LDA could calculate a distribution that took into account the commonality of scene descriptions and the overall content of the lyrics.

Future issues are as follows.

The first issue is to realize intelligent methods for selecting guide words. As far as the author experimented, depending on the guide words set, it sometimes happened that visualization results that reflected the contents of the lyrics could not be obtained. We would like to conduct further experiments to determine what kind of guide words are effective to obtain visualization results satisfactory to the user.

The second issue is the handling of lyrics containing multiple languages such as English. Our current implementation omits words in foreign languages and deals with only

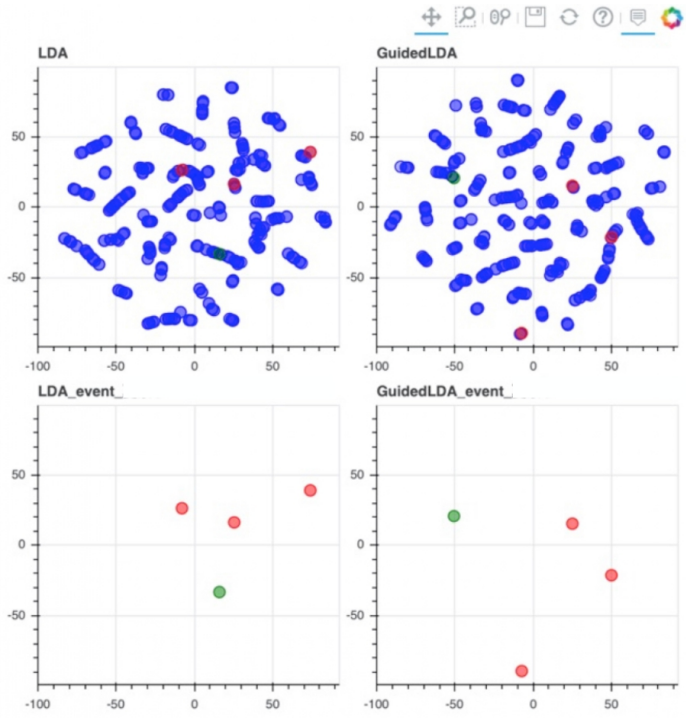

Fig. 5. (upper left) Application of LDA. (upper right) Application of guided LDA. (lower left) Only lyrics including ”sotsugyo” (graduation) and lyrics with content related to graduation (LDA). (lower right) Only lyrics including ”sotsugyo” (graduation) and lyrics with content related to graduation (guided LDA).

Japanese words. Since many recent J-POP songs include English phrases, we would like to include the semantics of English phrases in the visualization results. We need to discuss whether to translate them into Japanese lyrics, or whether to treat English words as separate words from Japanese words before extending our implementation.

The third issue is the color representation of visualization. It is effective to get an overview of the characteristics of the artists/lyricists and the trends of year or season by assigning the color of each point on the scatterplot corresponding to each song based on meta-information such as artist names, lyricist names, or released year or season. We would like to discover various correlations between lyrics and meta-information by implementing various color representations.

The fourth issue is the implementation of user interface functions. The goal of this study is to ”obtain visualization results that meet the user’s objectives by repeating a series of operations, such as inputting guide words and checking the visualization results,” but we have not yet implemented interactive mechanisms to realize this goal. We would like to extend the implementation so that we can interactively display the visualization results of the distribution of lyrics when a guide word is entered.

After extending our implementation based on the above issues, we would like to verify the visualization results with a larger number of songs and conduct user evaluation experiments.

REFERENCES

[1] Kobayashi, Y., Amagasa, M., and Suzuki, T. (2015). A Chronological Variation of Lexical Indices in the Lyrics of Popular Songs. Jinmonkon 2015, 2015, 23-30.

[2] Ohde, A., Matsumoto, A., and Kaneko, T. (2013). Lyrics changes of popular songs analyzed by age. Jinmonkon 2013, 2013(4), 103-110.

[3] Sadamura, K. (2019). Quantitative Text Analysis About Lyrics Written by Takashi Matsumoto. BULLETIN OF POLICY MANAGEMENT SHOBI UNIVERSITY, 34, 17-33.

[4] Kawamura. (2017). A Study on Music Recommendation System Considering Users’ Situation Based on Lyrics Information Analysis. Graduate School research report, Science and Engineering, 47.

[5] Hosoya, M., and Suzuki, T. (2010). Exploratory analysis of popular songs made by Japanese female singer-songwriters. Jinmonkon 2010, 2010(15), 195-202.

[6] Hossain, R., Sarker, M. R. K. R., Mimo, M., Al Marouf, A., and Pandey, B. (2019, February). Recommendation approach of english songs title based on latent dirichlet allocation applied on lyrics. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-4). IEEE.

[7] Sasaki, S., Yoshii, K., Nakano, T., Goto, M., and Morishima, S. (2014, October). LyricsRadar: A Lyrics Retrieval System Based on Latent Topics of Lyrics. In Ismir (pp. 585-590).

[8] Blei, D., Ng, A., and Jordan, M. (2001). Latent dirichlet allocation. Advances in neural information processing systems, 14.

[9] Jagarlamudi, J., Daumé III, H., and Udupa, R. (2012, April). Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 204-213)

[10] Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

[11] Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl 1), 5228-5235.