Most Recurring Topics in the Top 100 Billboard Songs

Vinícius R. M. Schmidt
Radboud University, Gelderland, The Netherlands
vinicius.ribeiromachadoschmidt@ru.nl

Music has been a widely adopted form of expression since the early civilizations. Even before music was accompanied by lyrics, it was a way of communicating feelings, emotions and themes. Once lyrics became popular, the expressed topic of songs became a bit easier to grasp. Given that, this project will focus on answering the question: Are certain topics more recurrent among the top popular songs in the U.S. since 1946 until today? Furthermore, I will also investigate whether the topics change throughout the years and if there is any relation between the topics and the songs' genres.

The dataset for this project consists of the lyrics to the top songs in the U.S. of each year since 1946, compiled by Billboard [2]. The songs are usually ranked by total sales (and total number of streams in recent years), but in some years they were ranked by how often they appeared in the top 100 weekly Billboard charts. Billboard only provides the top 100 songs from 2006 onwards on their official website, but all the Billboard top charts since 1946 can be found on Wikipedia [3]. This means that the song and artist names will have to be scraped from the Wikipedia web pages, which will then be used to query the Genius API [4] for the song lyrics.

When it comes to validation, I plan to use the perplexity measure used by [6] and the the UMass-coherence metric used in [5]. As an extra validation, I have found a website [1] which contains user annotated "moods and themes" for each song. The themes and moods listed on the website could be scraped for each song to validate the generated topics.

In terms of related work, other publications have covered similar topics in the past: Laoh et al. [2018] conducted topic modeling on Spotify's top 200 Indonesian songs between 2017 and 2018. The main challange here was analyzing lyrics written in Bahasa, considered to be a "complicated and complex language" [6]. Devi and Saharia [2020] performed topic modeling to classify sentiment from lyrics from 150 Manipuri songs. Finally, Siriket et al. [2021] used a combination of TFIDF and LDA to associate lyrics to moods and subsequently used that to power different machine learning classifiers.

REFERENCES

[1] ALLMUSIC, NETAKTION LLC 2024. *AllMusic.* ALLMUSIC, NETAKTION LLC. Retrieved October 10, 2024 from https://www.allmusic.com/

[2] Billboard Media, LLC. 2024. *Billboard.* Billboard Media, LLC. Retrieved October 9, 2024 from https://www.billboard.com/

[3] Wikimedia Foundation, Inc. 2024. *Billboard year-end top singles of 1946.* Wikimedia Foundation, Inc. Retrieved October 10, 2024 from https://en.wikipedia.org/wiki/Billboard_year-end_top_singles_of_1946

[4] ML Genius Holdings, LLC 2024. *Getting Started.* ML Genius Holdings, LLC. Retrieved October 10, 2024 from https://docs.genius.com/

[5] Maibam Debina Devi and Navanath Saharia. 2020. Exploiting Topic Modelling to Classify Sentiment from Lyrics. In *Machine Learning, Image Processing, Network Security and Data Sciences*, Arup Bhattacharjee, Samir Kr. Borgohain, Badal Soni, Gyanendra Verma, and Xiao-Zhi Gao (Eds.). Springer Singapore, Singapore, 411–423. https://doi.org/10.1007/978-981-15-6318-8_34

[6] Enrico Laoh, Isti Surjandari, and Limisgy Ramadhina Febirautami. 2018. Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE).* IEEE, 270–274. https://doi.org/10.1109/ICISCE.2018.00064

[7] Kornkanya Siriket, Vera Sa-ing, and Subhorn Khonthapagdee. 2021. Mood classification from Song Lyric using Machine Learning. In *2021 9th International Electrical Engineering Congress (iEECON).* IEEE, 476–478. https://doi.org/10.1109/iEECON51072.2021.9440333

Conference'17, July 2017, Washington, DC, USA