Project Proposal: NLP techniques to analyze song lyrics

ABHINAV GHAI and ALEX TANASESCU

ACM Reference Format:
Abhinav Ghai and Alex Tanasescu. 2023. Project Proposal: NLP techniques to analyze song lyrics. 1, 1 (January 2023), 3 pages.

1 INTRODUCTION

This project will be an exploration into analysis of music lyrics. The primary goal will be to apply topic modelling techniques on lyrics to identify unique words or phrases that are most identifiable with a specific theme. There will also be an analysis of the lyrics in an attempt to place them into genres. Secondly, the project will look into document embedding techniques, such as doc2vec, that will allow us to explore the similarity between song vector representations of different artists, and attempt to see if we can produce an artist-song embedding, by "averaging" the lyric vectors, and use it to predict if a set of lyrics belongs to that artist, or similarly using a genre-song embedding to predict if a song belongs to a specific genre. We expect that lyrics won't play such a large part in the genre of a song, but will for song lyrics from a specific artist.

2 RELATED WORK

Analyzing song lyrics for a specific artist in many similar works starts off by first preprocessing the lyrics to create an updated stop words list. After filtering them out, and conducting other preprocessing techniques such as lemmatization, feature extraction techniques are typically applied. In particular, Tf-idf and bag of words can be utilized to analyze these lyrics, however results through this were sometimes not accurate enough, as they served primarily as a way to normalize the data, but they didn't consider the context provided by surrounding words in the lyrics, which could heavily warp the results they produced (and thus leading to inaccurate topic models when a topic modelling algorithm was applied to the results). As a result, most previous works supplemented this with a vector representation for lyrics. One way that usually emerged was the use of Word2vec. Word2vec allows for the context surrounding a specific word to hold greater weight. Similarly for the document embedding portion of the project, there exist previous works which have analyzed lyrics in great depth using doc2vec, they follow the same general process, the major difference being that a document embedding is produced, instead of a word embedding, this allows the analysis of a song as an individual entity, compared to the previously mentioned Word2vec and tf-idf analysis methods which more-so concerned themselves with individual words (and their associated contexts). Furthermore, most related work displayed these vector results using a technique called t-Distributed Stochasitc Neighbour Embedding (t-SNE) which allowed them "convert" these multi-dimensional vectors into 2d vectors to allow them to be more easily visualised, this is likely a technique our project will utilize as well, due to its ability to convert the results into a more intuitive format. Similarity analysis is also a minor topic our project delves into, previous works have utilized cosine similarity analysis, which is

Authors' address: Abhinav Ghai; Alex Tanasescu.

Manuscript submitted to ACM

one of the many ways to compare similarity between two vectors, this was utilized to compare different a lyric to a lyric-artist embedding typically to evaluate the results.

3 PROPOSED WORK

We plan on using the following datasets for our studies as they include various artists that allow for diversity among themes, genres, artists and most importantly song lyrics:

- https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset
- https://www.kaggle.com/datasets/nikhilnayak123/5-million-song-lyrics-dataset
- https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres?resource=download

3.1 Topic Modelling Common Themes

For the first part of our project consists of analyzing songs from artists using the dataset mentioned above. After we preprocess the data for each artist we are going to be using gensim to conduct LDA on the data. Finally, we plan on manually picking the top topic names for the artists, based on the results of the LDA, and presenting them visually through t-SNE.

3.2 Comparisons between document embeddings

For the second part of our project we will continue using the above data sets to extract our data. However, this time, we will analyze the top 10 most common genres in the datasets and use gensim to produce document embeddings for songs in those genre, we will then average them to produce the "genre-lyric" embedding, and then use cosine similarity to evaluate how close songs from the same genre are, and evaluate how well we can determine if a certain lyrics are part of a specific genre. The document embeddings for genres will also be presented in a visual format with the help of t-SNE.

[7], [6], [5], [3], [8], [15], [4], [12], [10], [9], [2], [1], [11], [14], [13],

REFERENCES

[1] Jia Yi Chan. 2021. *How to do AVERAGE and MAX word embedding for LONG sentences?* Retrieved January 29, 2023 from https://towardsdatascience.com/how-to-do-average-and-max-word-embedding-for-long-sentences-f3531e99d998

[2] Timothy James Dobbins. 2018. *Analyzing Rap Lyrics Using Word Vectors.* Retrieved January 29, 2023 from https://tmthyjames.github.io/2018/january/Analyzing-Rap-Lyrics-Using-Word-Vectors/

[3] Ian Freed. 2018. *Using Machine Learning to Analyze Taylor Swift's Lyrics.* Retrieved January 29, 2023 from https://www.codecademy.com/resources/blog/taylor-swift-lyrics-machine-learning/

[4] Dilyan Kovachev. 2019. *How We Used NLTK and NLP to Predict a Song's Genre From Its Lyrics.* Retrieved January 29, 2023 from https://towardsdatascience.com/how-we-used-nltk-and-nlp-to-predict-a-songs-genre-from-its-lyrics-54e338ded537

[5] Susan Li. 2018. *Multi-Class Text Classification with Doc2Vec Logistic Regression.* Retrieved January 29, 2023 from https://towardsdatascience.com/multi-class-text-classification-with-doc2vec-logistic-regression-9da9947b43f4#:~:text=Doc2vec%20is%20an%20NLP%20tool,of%20scope%20of%20this%20article.

[6] Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2020. *Document Representation.* Springer Singapore, Singapore, 91–123. https://doi.org/10.1007/978-981-15-5573-2_5

[7] Shay Palachy. 2019. *Document Embedding Techniques.* Retrieved January 29, 2023 from https://towardsdatascience.com/document-embedding-techniques-fed3e7a6a25d#a2f6

[8] Federico Pascual. 2019. *Document Embedding Techniques.* Retrieved January 29, 2023 from https://monkeylearn.com/blog/introduction-to-topic-modeling/

[9] Adam Reevesman. 2020. *Lyric-based song recommendation with Doc2Vec embeddings and Spotify's API.* Retrieved January 29, 2023 from https://towardsdatascience.com/lyric-based-song-recommendation-with-doc2vec-embeddings-and-spotifys-api-5a61c39f1ce2

[10] Alda Sianipar. 2019. *Predicting a Song's Genre Using Natural Language Processing.* Retrieved January 29, 2023 from https://betterprogramming.pub/predicting-a-songs-genre-using-natural-language-processing-7b354ed5bd80

[11] tsandefer. 2019. Doc2Vec and Annotated Lyrics: Are they "Genius?". https://github.com/tsandefer/dsi_capstone_2

[12] Dea Venditama. 2020. *Music Lyrics Analysis Using Natural Language Processing.* Retrieved January 29, 2023 from https://medium.com/analytics-vidhya/music-lyrics-analysis-using-natural-language-processing-7647922241c0

[13] Wikipedia. 2023. Cosine similarity. https://en.wikipedia.org/wiki/Cosine_similarity

[14] Wikipedia. 2023. Similarity measure. https://en.wikipedia.org/wiki/Similarity_measure

[15] Joyce Xu. 2018. *Topic Modeling with LSA, PLSA, LDA lda2Vec.* Retrieved January 29, 2023 from https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05