**Data**

We collected data from Mendeley Data, which provides a list of lyrics from 1950 to 2019 describing music metadata as sadness, danceability, loudness, acousticness, etc., and also other information including artist name, music genre, and so on [1]. We selected both genre and lyrics from the original dataset and did natural language processing to clean the data. We deleted all non-English words in the lyrics, removed punctuation and double spaces, and changed all words to lower case in order to better encode them in our models. Since words can be grouped together based on their lemma and be analyzed as a single item or token, we lemmatized words in lyrics using the WordNet's built-in morphy function. In addition, words such as ‘the’, ‘a’, ‘in’ provide no useful information in lyric classification. Therefore, we used ‘stopwords’ packages from the NLTK library to remove all stop words in English. After data pre-processing, our final dataset contains 28,031 lyrics in seven genres: blues, country, hip hop, jazz, pop, reggae, and rock. For model training and testing, we split the training and testing data sets in an 80/20 ratio.

**Generative Model**

Describe the model in detail and develop a solution using parameter inference (and/or decoding).

**Apply to Real Data:** Highlight situations where each approach performs well and poorly.

We chose Latent Dirichlet Allocation (LDA) as our generative model and used the Latent Dirichlet Allocation with the online variational Bayes algorithm package from scikit-learn [2][3] to classify seven music genres. Since LDA is a type of unsupervised machine learning, we can only specify how many topics we want to find, which is seven. The seven classified topics that were given by the model were not necessarily the same as our original genres. After fitting the data into the model, we printed the top 50 words in each topic and used the pyLDAvis package to visualize the model results [4]. Table 1 shows the top 50 words in each topic. As shown in Figure 1, there are some overlaps between Topic1 and Topic 2. Since for each document there is a topic distribution, we then used the argmax function on each document’s topic distribution to assign each document to a topic. In order to map the topics that were modeled by LDA to our original genres, we found the top 50 words in each original genre, computed the number of overlap words between each genre and each topic, and mapped each topic to the genre with the most overlapping words.

***Table 1:*** *Top 50 Words in Each Topic*

|  |  |
| --- | --- |
| **Topic** | **Top 50 Words** |
| 0 | world, like, bitch, time, real, high, smoke, fuck, people, yeah, know, life, need, face, live, cause, come, hell, roll, money, look, place, tell, city, stay, sell, think, go, watch, teach, right, work, shit, fake, young, ghost, deal, line, ride, stick, rule, paper, read, straight, race, drug, drink, mind, school, feel |
| 1 | heart, baby, go, hold, know, long, night, tonight, leave, right, believe, like, time, want, kiss, stay, come, feel, need, gonna, break, eye, wait, tear, tell, cause, arm, girl, sweet, hand, close, little, yeah, darling, love, start, touch, look, promise, apart, say, wrong, wanna, true, think, dream, fall, tight, light, line |
| 2 | away, good, break, walk, know, lonely, feel, night, go, wish, fool, miss, stand, come, dream, leave, goodbye, gonna, yeah, tell, morning, heart, cry, sleep, little, look, whoa, cause, fade, rain, think, tear, say, like, take, baby, girl, time, memory, hurt, want, stay, blue, right, place, wonder, home, dear, hide, yesterday |
| 3 | fall, feel, come, eye, know, like, inside, head, cold, black, dead, blood, hand, stand, soul, burn, hear, fear, leave, pain, fight, face, save, lose, wall, kill, death, devil, lie, turn, tear, light, grind, die, wind, hide, follow, look, break, mind, word, heart, deep, hold, live, hell, speak, life, body, dream |
| 4 | like, yeah, know, cause, fuck, come, shit, better, wanna, want, money, gonna, tell, gotta, play, right, party, bout, shoot, think, girl, talk, stop, say, head, look, bitch, little, need, game, start, baby, feel, damn, good, kick, drop, get, beat, go, hand, everybody, watch, check, goin, work, black, pull, night, turn |
| 5 | sing, home, song, blue, hear, come, woman, play, music, right, sweet, remember, go, lord, bring, like, ring, long, hand, word, songs, listen, write, call, know, money, say, fight, land, sound, yeah, days, little, radio, night, people, band, time, tell, dance, guitar, summer, swing, head, christmas, kill, year, roll, soon, tune |
| 6 | time, life, live, know, change, mind, come, think, want, world, things, need, cause, dream, feel, like, lose, love, look, go, today, give, yeah, tell, days, hard, right, learn, better, leave, forget, true, gonna, wanna, hurt, somebody, take, tomorrow, turn, remember, make, long, reason, forever, inside, thing, ready, matter, easy, mean |

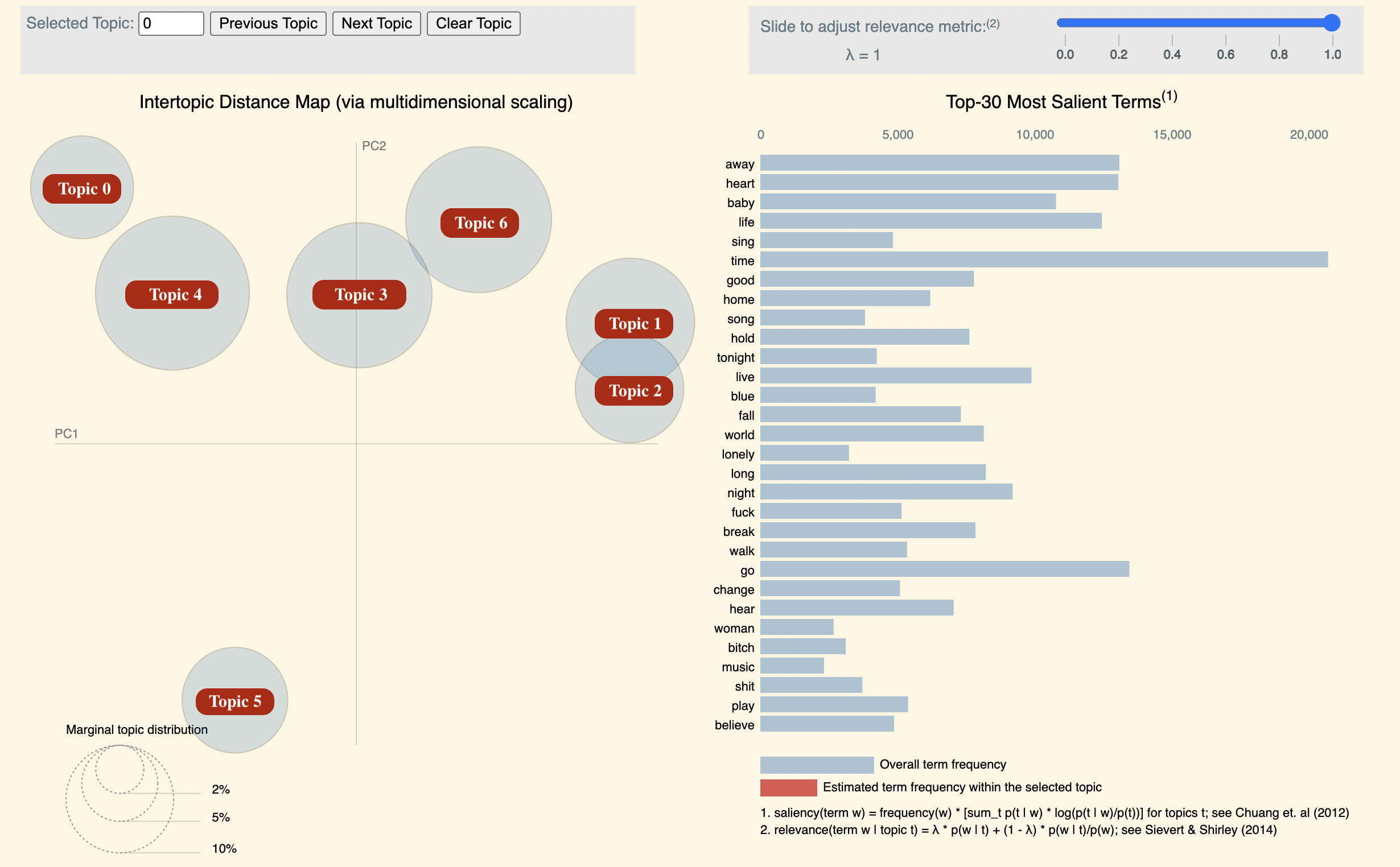
***Figure 1:*** *Visualization using pyLDAvis*

Table 2 shows the mapping results of topics to genres as well as the result of applying our LDA model to real data, which achieved an overall correctness of 24.02% with relatively higher correctness on rock and hip-hop and relatively lower correctness on jazz and reggae.

***Table 2:*** *Mapping Topics to Genre & Real Data Results*

|  |  |  |
| --- | --- | --- |
| **Topic** | **Genre** | **Correctness** |
| 0 | Jazz | 0.0602 |
| 1 | Pop | 0.1849 |
| 2 | Country | 0.1361 |
| 3 | Rock | 0.3363 |
| 4 | Hip-hop | 0.6737 |
| 5 | Reggae | 0.0977 |
| 6 | Blues | 0.1926 |

**Synthetic Data**

Evaluate the results qualitatively and quantitatively. Highlight situations where each approach performs well and poorly.

We used the LDA model that we trained to generate synthetic data. For each genre, we generated 1,000 lyrics. The length of each lyric is randomly chosen from the length of lyrics in the real data[5], and the words in each lyric are chosen based on the distribution of words for each topic.

Table 3 shows the samples of generated data for each genre. Table 4 shows the results of applying our LDA model to the synthetic data that is generated by itself. Since the generated data meets all the assumptions of our LDA model, our model achieved correctness above 99% across all genres.

***Table 3:*** *Synthetic Lyrics for Each Genre*

|  |  |  |
| --- | --- | --- |
| **Topic** | **Genre** | **Synthetic Lyrics** |
| 0 | Jazz | blueprint sponsor daily see orange breed need jones world death bout unlikely scream play higher arguments abhorrent sense smoke dust hustle slide deal smoke high staffed struggle beast beat read world tick dodge people |
| 1 | Pop | tonight know hold lover free care stay crime long letter need long hold long break middle baby keep till heart long go kiss hold night darling heart right lover perfect ring heart sweet right stay little hold night hold mistake cause picture |
| 2 | Country | stocking news blow sleep cry away blue lovely leave break say dolls suddenly sweet tell cry piece turn lonely need farewell say come alright give |
| 3 | Rock | blind recover warm crack cruel world throne know roses mother greatest near come higher save funeral defect fall mother mouth idea line shape piece room doctor fight see shake feel shiny wire |
| 4 | Hip-hop | cause bout body make card fuck say gotta white know bitch everybody wanna gotta count beat buckle niki start shift right patch mama cold fuck tip sure noise fireworks turn better music clean thing ashamed hear wild seat everyday bloody |
| 5 | Reggae | sweet home good home come smile belong sing balloon go boys bloody kings music promise dine kill friend river brother little high look inspiration woman understand home sweet happen hand |
| 6 | Blues | hug break kind hide life everyday go keep shame know single time life late feel free thoughts destination world life believe gonna better room thing hate loan sure lose yeah outside dice time world dream honestly weasel mean better live world |

***Table 4:*** *Synthetic Data Results*

|  |  |  |
| --- | --- | --- |
| **Topic** | **Genre** | **Correctness** |
| 0 | Jazz | 0.9987 |
| 1 | Pop | 0.9987 |
| 2 | Country | 0.9976 |
| 3 | Rock | 1.0000 |
| 4 | Hip-hop | 0.9988 |
| 5 | Reggae | 0.9987 |
| 6 | Blues | 0.9988 |

**Discussion**

Discuss pros and cons of the two approaches. Consider:

• quality/correctness  
• data, time, and computational requirements

• interpretability

The "bag of-words" assumption of the LDA model lost information about the order of the words when trying to classify lyrics. Also, the LDA model performs worse at capturing rare topics due to data quality issues. Due to an increase in observations, our LDA model can be improved by training with a larger dataset. With the LDA, documents are represented as mixtures over latent topics, but when we assigned topic labels to lyrics, we simplified by choosing the topic with the highest probability, and mapping the pre-annotated genre of each lyric to a topic based on the number of overlapping words. A more reasonable approach can be taken to further improve the model's performance. During the process of generating data, since we randomly chose words based on the distribution of words for each topic, as we can see from Table 3, the synthesized lyrics are simply accumulations of words instead of meaningful sentences.

Table 5 shows the result of the two models on both the real and synthetic data. Compared to our discriminative model, the LDA model performed worse on the real data but better on the synthetic data that was generated by itself.

***Table 5:*** *Overall Correctness*

|  |  |  |
| --- | --- | --- |
| **Model** | **Correctness**  **(Real Data)** | **Correctness**  **(Synthetic Data)** |
| LDA | 0.2402 | 0.9988 |
| BERT |  |  |

The LDA model has fewer computational requirements as it can be trained locally within 10 seconds. With the algorithm and related distribution of a LDA model, it is more interpretable.

**Reference**

[1] L. Moura, E. Fontelles, V. Sampaio, and M. França, “Music Dataset: Lyrics and Metadata from 1950 to 2019,” *data.mendeley.com*, vol. 2, Aug. 2020, doi: 10.17632/3t9vbwxgr5.2.

[2]“blei-lab/onlineldavb,” *GitHub*, May 17, 2022. https://github.com/blei-lab/onlineldavb

[3] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic Variational Inference,” *Journal of Machine Learning Research*, vol. 14, no. 4, pp. 1303–1347, 2013, Accessed: May 18, 2022. [Online]. Available: https://jmlr.org/papers/v14/hoffman13a.html

[4] B. Mabey, “bmabey/pyLDAvis,” *GitHub*, Oct. 15, 2020. https://github.com/bmabey/pyLDAvis

[5] A. Dai, “Duke NLP Final Project,” *GitHub*, Jan. 02, 2022. https://github.com/dai-anna/Duke-NLP-FinalProject

‌