An implementation and Evaluation of a Topic Model for Song Lyrics

Introduction and Task Motivation

Music is cool, Music is really cool. The sphere of music is massive, regarding both its diversity and sheer quantity of content. Millions upon millions of songs come together to produce a variety of sounds, moods, genres, tempos, levels of danceability, and more. One thing about music, however, is that it can only be so standardized, as part of the beauty of music is how open to interpretation it is for everyone. In recent years Spotify has begun to implement AI into its applications like the DJ feature, which is supposed to act as your personal DJ, and AI generated playlists. As users of Spotify and other music streaming apps, we are very aware of the heart, soul, and nuance that can go into curating playlists. It is truly a skill, and at the core of that skill is the ability to experience a song and attribute it to some topic, genre, or mood. Despite advances in text analysis, identifying emotional nuances in creative works remains a challenge due to their inherent subjectivity.

With that being said, we wanted to see how well the human thought process behind categorizing songs could be conveyed by a topic model in comparison to both our own and that of Chatgpt. We created an LDA topic model using the Gensim library and trained one model on a dataset composed of song lyrics. We then created duplicate versions of the model with a different number of passes to emulate relistening to music, looking to see whether topics and evaluation scores improved. We then fed the model 45 songs to match to its generated topics, asked ChatGPT to match the 45 songs to the topics generated by our topic model implementation, and finally did the same things ourselves. We want to learn and evaluate how capable and subjectively accurate topic models can be in categorizing songs, and discuss the possible shortcomings that could come from the model only having access to the lyrics.

The motivation for this project comes from curiosity and love for music. A lot of us have come across all kinds of new music thanks to these streaming apps' ability to recommend them to us. This project presented us an opportunity to explore that at a naive level.

Description of dataset including context of its content along with descriptive statistics of its dimensions and how it was collected

The dataset selected to train the model for this project is this Spotify Million Song Dataset which was found on Kaggle. The specifics of how the data in this dataset was obtained is unclear, but we believe this was done via a web scraper or a series of api calls by the curator. The dataset contains 57494 different song lyrics, all accompanied by the song name, 643 unique artists, and a link to the song. This dataset is fitting for this project because it is not excessively large, making it manageable to run code on, while still including modern songs from a contemporary streaming service. This makes the project and its findings more relevant in the sense that they stand up to time in this moment.

The dataset used to evaluate/test the model is a set of 45 songs curated by us, with 15 coming from each member of the group. These are songs that we listen to and can be found here. This dataset contains the name of the song and the lyrics. We put the songs together into a playlist and manually copy pasted their lyrics into the dataset.

Evaluation Metrics

In this project, we used three different points of comparisons to evaluate the performance of our topic model: The categorization of the different topics by our topic model implementation, the categorization of the different song lyrics by ChatGPT and the human categorization of the song lyrics. To evaluate the performance of our topic model implementation, we used the following metrics: Perplexity, Coherence scores (U_Mass and U_uci), topic diversity, topic consistency and topic interpretability.

The perplexity evaluation metric enables us to gauge how well the topic model performs with unseen data. Generally, a lower perplexity score indicates a better performance of the model, with values in the range [-4, 0] indicating a better fit, [-7, -4] indicating a reasonably good model and values <-10 indicating overfitting. The U_Mass coherence score-typically ranging from 0 to negative infinity- calculates how often two words appear together in a corpus, while the U_uci coherence score calculates how often these two words appear together in a given window. Practically, a higher coherence score indicates that words within the same topic appear together in songs, translating better semantic similarity between the top words in a given topic, hence suggesting better interpretability. The U_uci coherence metric goes further than the U_Mass to consider how close the given words appear together in the corpus, which makes it more sensitive to the word order in song lyrics, and consequently better at capturing semantic relationships within a given topic.

To understand how unique our topics are relative to each other, we implemented a topic diversity metric. The diversity score is normalized to give values in the range [0,1]. This metric enables us to gauge how different words overlap between the various topics that the model generates, with lower scores indicating more overlapping topics, and greater scores indicating more distinct topics. This is reflected in the visualization, where the models with higher topic diversity values have less circles that overlap with each other, whereas the topic models with lower diversity values have more circles that overlap with each other. This metric gives us an idea of how well the model performs at capturing a variety of topics found in our dataset of song lyrics.

In addition to seeing how well our model catches diverse topics, we were interested in understanding the degree to which the different categorizations are able to stay focused on a single, relevant topic. By implementing the topic consistency metric, we are able to gauge how well the words and phrases of the songs that appear in a given topic consistently relate to the same central theme, without significant deviations to unrelated subjects. Low topic consistency scores would show that similar songs might get different topic distribution weights, while higher

topic consistency scores indicate that similar songs get similar topic distribution weights. For instance, consider doc1 and doc2 to be similar songs about love. A model with a low topic consistency value might distribute doc1 as 40% topic1(let's say love) and 60% topic2(let's say party), and distribute doc2 as 70% topic1 and 30% topic2. The model being inconsistent in how it classifies content which are similar suggests it is less confident. On the other hand, a model with higher topic consistency values might distribute the same doc1 as 75% topic1 and 25% topic2, and doc2 as 80% topic1 and 20% topic2, showing a greater consistency in the classification, and hence a greater confidence of the model about what songs belong where.

Finally, it is important for us humans using the model to easily understand and explain the meaning of the topics spotted by the topic modeling algorithm, so we use the topic interpretability metric. The values are in the rang [0,1], with higher scores indicating the words in the topic are more semantically related and hence the topic is more interpretable by humans, while a low interpretability score would mean the words shown by the model are less semantically related, and the topic more difficult to interpret from a human perspective.

Our topic model implementation was run with a varying number of passes, the evaluation metrics recorded for every different number of passes and the results are presented in a table in the result section. To get a more tangible idea of the performance of the topic model, we carried out a human evaluation, where given the classification produced by the topic model, we asked three different people the questions: How well do the words in each topic fit together?; How easy is it to assign labels?; How well do the topics represent a variety of music categories? We equally asked them to assign labels to the different topics based on the words that featured, and compared the different labels that different people assigned to the various topics. This helps give another evaluation criteria on how well the topics generated agree with human interpretation, and provide a comparison with how ChatGPT interprets them as well.

We will also be evaluating how well the topic attribution for the songs in the testing dataset align across our model, ChatGPT, and us.

Baseline approach and Results

Our baseline approach is building an LDA topic model, using the Gensim library, trained on the Spotify dataset. After tokenizing the dataset, we applied filtering to remove very common words like 'I' and 'the'. We then used Gensim’s LDA model to generate topics from the Spotify song lyrics dataset. We used a word cloud to visualize the topics by displaying the top 10 words from each topic. The size of each word represents its contribution to the topic, highlighting the differences in significance.

Fig: Word cloud representing topic model for 20 topics on 10 passes

Some topics are clearer than others in terms of the story/tone they portray like topics 8 and 6, which seem to portray stories of love/romance and murder/aggression respectively. Songs and creative works can be really niche and specific. Considering that there are around 200 to 500 different genres for English music existing, which could tell stories of literally anything, we would say the topics generated by the model are somewhat respectable. Part of why topic modeling seemed so compatible with something as subjective as music is because the process for topic models is unsupervised and its output is just as subjective. All in all, we were pretty impressed with the topics that the model generated.

However, we must be understanding of the model's limitations, as it cannot perceive music in the same way humans do. Nonetheless, we were still curious as to what it heard and what kind of sound might be produced by these topics. Luckily, we discovered this Spotify API that has created audio feature statistics for each of the songs in their database. So, we utilized this API to evaluate the presence of criterions like danceability, tempo and loudness among the topics generated by our topic model.

	topic	avg_danceability	avg_energy	avg_key	avg_loudness	avg_mode	avg_speechiness	avg_instrumentalness	avg_liveliness	avg_valence	avg_tempo
0	0	0.50	0.57	4.56	-9.14	0.79	0.04	0.05	0.21	0.40	121.56
1	1	0.55	0.69	5.12	-8.08	0.73	0.06	0.07	0.23	0.59	122.89
2	2	0.52	0.57	4.17	-9.18	0.75	0.06	0.10	0.25	0.58	128.32
3	3	0.58	0.78	5.78	-7.78	0.56	0.05	0.10	0.21	0.75	112.32
4	4	0.54	0.65	6.17	-9.43	0.83	0.05	0.07	0.22	0.54	124.15
5	5	0.56	0.55	5.15	-10.17	0.77	0.11	0.13	0.25	0.69	137.66
6	6	0.57	0.76	4.09	-7.87	0.64	0.08	0.22	0.25	0.46	111.56
7	7	0.57	0.67	6.56	-9.55	0.78	0.05	0.01	0.22	0.68	120.38
8	8	0.57	0.58	5.07	-10.26	0.80	0.06	0.10	0.24	0.66	124.58
9	9	0.52	0.64	7.18	-10.20	0.82	0.04	0.02	0.14	0.53	125.23
10	10	0.41	0.75	8.00	-6.93	0.75	0.05	0.22	0.60	0.56	148.30
11	11	0.58	0.62	4.11	-8.89	0.89	0.04	0.02	0.28	0.62	128.05
12	12	0.59	0.81	4.50	-6.85	0.92	0.05	0.06	0.16	0.84	142.34
13	13	0.48	0.79	4.75	-6.34	0.62	0.07	0.11	0.26	0.40	127.08
14	14	0.56	0.66	5.27	-7.65	0.73	0.09	0.01	0.27	0.56	128.51
15	15	0.58	0.60	5.83	-11.64	0.50	0.08	0.07	0.33	0.59	129.21
16	16	0.52	0.53	5.32	-9.91	0.93	0.05	0.04	0.29	0.38	116.71
17	17	0.53	0.69	5.74	-8.35	0.59	0.06	0.10	0.20	0.47	124.48
18	18	0.51	0.50	4.43	-11.13	0.86	0.06	0.04	0.19	0.47	112.92
19	19	0.58	0.75	5.12	-8.78	0.62	0.07	0.01	0.23	0.60	125.10

We think this data shows even more that the interpretation of music only goes so far without that human side of it. We expected some of our sadder topics to have lower levels of energy considering what most sad songs that come to one's mind immediately sound like, but that wasn't the case. Looking at this table might lead one to believe that the songs were pretty similar, in reality we don't think they were. The thing is, some of the happiest sounding songs could be telling the saddest stories, and that adds an added layer of complexity as a non-human recommender. You could suggest a happy sounding sad song to a person building a sad playlist and technically be right, but it also wouldn't be crazy to see that user displeased by that recommendation. However, it could also be a reflection of how trendy the music industry is, and how a lot of people talk about some of the same themes in their songs. We also must take all of these statistics with a grain of salt. Although the logic behind some of these stats is available on their web api's documentation, we found that the labels can be considered misleading. Topic 10 having the lowest danceability is expected, as we imagine that a lot of songs fitting into that theme are more storytelling based and talk to the listener like country songs. Songs like this would lack the structure to make it danceable, but there's a layer of nuance as to what kind of dance they're referring to.

All in all, this table could make us think that the model was incapable of hearing anything really when making these categories and that the lyrics just can't be enough to produce an accurate representation of the song. On the other hand, we could regard it as the model being able to hear everything it needs to, as songs of all themes can come in all shapes and sizes. With that being said, topic modeling shows itself to only be so ideal for things like song categorization and understanding.

Third-party tool and Results

As previously mentioned, we utilized ChatGPT alongside human input to analyze and compare the differences in how the songs in our curated playlist were categorized into topics. To do this we prompted ChatGPT with our 20 unnamed topics in the form of a list where each element was a tuple of the topic and each of its 10 words pertaining to that topic and their weights.

[(0,

'0.110*"life" + 0.081*"world" + 0.070*"live" + 0.062*"dream" + 0.049*"believe" + 0.033*"free" + 0.031*"find" + 0.025*"hope" + 0.020*"end" + 0.016*"alive"'),

(1,

'0.073*"time" + 0.052*"way" + 0.047*"tell" + 0.044*"think" + 0.035*"thing" + 0.032*"leave" + 0.030*"try" + 0.026*"say" + 0.025*"mind" + 0.022*"find"'),

(2,

'0.253*"day" + 0.069*"blue" + 0.063*"new" + 0.059*"stop" + 0.046*"rain" + 0.046*"lonely" + 0.044*"year" + 0.040*"happy" + 0.022*"sad" + 0.019*"thank"'),

(3,

'0.208*"good" + 0.196*"girl" + 0.174*"little" + 0.048*"crazy" + 0.036*"fine" + 0.030*"lady" + 0.024*"bit" + 0.021*"guy" + 0.019*"babe" + 0.019*"pretty"'),

(4,

'0.043*"light" + 0.028*"sun" + 0.027*"fly" + 0.026*"sky" + 0.022*"shine" + 0.021*"star" + 0.018*"wind" + 0.013*"water" + 0.013*"moon" + 0.012*"sea"'),

(5,

'0.049*"old" + 0.030*"ride" + 0.025*"say" + 0.025*"town" + 0.022*"drive" + 0.019*"drink" + 0.019*"car" + 0.017*"house" + 0.014*"kid" + 0.013*"daddy"'),

(6,

'0.022*"fuck" + 0.022*"shit" + 0.020*"hit" + 0.017*"bitch" + 0.014*"kill" + 0.012*"drop" + 0.010*"hate" + 0.010*"shoot" + 0.008*"check" + 0.008*"ass"'),

(7,

'0.215*"man" + 0.118*"run" + 0.060*"woman" + 0.056*"child" + 0.043*"hand" + 0.041*"young" + 0.038*"bear" + 0.031*"help" + 0.025*"son" + 0.024*"mother"'),

(8,

'0.557*"love" + 0.100*"need" + 0.042*"kiss" + 0.037*"sweet" + 0.035*"true" + 0.033*"heart" + 0.020*"lover" + 0.011*"darling" + 0.008*"lip" + 0.007*"arm"'),

(9,

'0.376*"baby" + 0.105*"dance" + 0.077*"rock" + 0.063*"roll" + 0.057*"shake" + 0.040*"music" + 0.032*"honey" + 0.019*"body" + 0.014*"rhythm" + 0.010*"sugar"'),

(10,

'0.463*"go" + 0.072*"long" + 0.068*"home" + 0.043*"road" + 0.038*"round" + 0.021*"mile" + 0.019*"worry" + 0.018*"carry" + 0.010*"highway" + 0.010*"wheel"'),

(11,

'0.143*"wanna" + 0.048*"wish" + 0.042*"beat" + 0.035*"eat" + 0.019*"shout" + 0.016*"ice" + 0.013*"minute" + 0.013*"card" + 0.013*"action" + 0.012*"freak"'),

(12,

...

'0.080*"look" + 0.059*"eye" + 0.040*"cry" + 0.038*"stand" + 0.037*"die" + 0.037*"walk" + 0.037*"see" + 0.036*"face" + 0.031*"head" + 0.022*"watch"'),

(18,

'0.171*"come" + 0.121*"let" + 0.110*"feel" + 0.073*"night" + 0.053*"hold" + 0.041*"wait" + 0.020*"close" + 0.020*"open" + 0.019*"remember" + 0.018*"touch"'),

(19,

'0.332*"get" + 0.066*"be" + 0.042*"big" + 0.041*"money" + 0.032*"work" + 0.022*"got" + 0.021*"lot" + 0.019*"pay" + 0.013*"city" + 0.011*"buy"')]

We then fed ChatGPT the song name and lyrics and asked it to place it in one of the 20 topics. For the human input we looked at all the topic word clouds and compared it to each of the song lyrics and selected which topic we felt correlated the closest with the topic. Looking at the results:

	song_name	our_placement	chatgpt_placement	model_dominant_topic	topic_probability
0	No One Noticed	Topic 2	Topic 1	Topic 1	0.20
1	Don't Stop The Music	Topic 9	Topic 9	Topic 1	0.15
2	Palms	Topic 10	Topic 18	Topic 0	0.26
3	Rude	Topic 8	Topic 10	Topic 19	0.17
4	Poison Ivy	Topic 3	Topic 6	Topic 13	0.20
5	Hellraiser	Topic 9	Topic 6	Topic 0	0.20
6	Cruisin' for a Bruisin	Topic 5	Topic 9	Topic 5	0.22
7	In the End	Topic 13	Topic 0	Topic 1	0.37
8	Unwritten	Topic 0	Topic 0	Topic 18	0.21
9	Sprinter	Topic 19	Topic 7	Topic 6	0.16
10	Crooked Smile	Topic 3	Topic 7	Topic 1	0.23
11	Get You	Topic 8	Topic 8	Topic 1	0.20
12	Ode To The Mets	Topic 17	Topic 2	Topic 1	0.22
13	Workin Out	Topic 13	Topic 0	Topic 6	0.23
14	E. Coli	Topic 0	Topic 6	Topic 0	0.12
15	Snooze	Topic 8	Topic 8	Topic 0	0.19
16	Survive (Ban Rap)	Topic 13	Topic 6	Topic 0	0.20
17	Broken Vessels (Amazing Grace)	Topic 17	Topic 0	Topic 17	0.35
18	Last Time I Saw You	Topic 13	Topic 1	Topic 1	0.26
19	Motivation	Topic 15	Topic 1	Topic 1	0.26
20	Centuries	Topic 1	Topic 5	Topic 18	0.17
21	Courtesy Call	Topic 16	Topic 6	Topic 1	0.23
22	Best Mistake	Topic 8	Topic 8	Topic 1	0.23
23	Home	Topic 10	Topic 10	Topic 1	0.28
24	Send My Love (To Your New Lover)	Topic 8	Topic 8	Topic 8	0.19
25	Legends Never Die	Topic 12	Topic 5	Topic 17	0.28
26	How To Love	Topic 8	Topic 8	Topic 8	0.17
27	Boo'd Up	Topic 9	Topic 8	Topic 1	0.27
28	Adore You	Topic 3	Topic 8	Topic 8	0.28
29	Hold On	Topic 18	Topic 0	Topic 18	0.21
30	Mary On A Cross	Topic 10	Topic 7	Topic 1	0.16
31	Them Changes	Topic 13	Topic 8	Topic 13	0.22
32	The Color Violet	Topic 13	Topic 8	Topic 9	0.24
33	Rockstar	Topic 19	Topic 6	Topic 19	0.14
34	Fishin' in the Dark	Topic 4	Topic 5	Topic 4	0.20
35	3 Nights	Topic 14	Topic 10	Topic 1	0.23
36	End of Beginning	Topic 18	Topic 0	Topic 0	0.20
37	Hotel California	Topic 8	Topic 7	Topic 17	0.14
38	Southern Nights	Topic 4	Topic 10	Topic 0	0.14
39	Birds of a Feather	Topic 17	Topic 4	Topic 1	0.23
40	Play That Funky Music	Topic 9	Topic 9	Topic 9	0.27
41	Smooth Operator	Topic 7	Topic 9	Topic 6	0.38
42	Higher	Topic 0	Topic 0	Topic 1	0.22
43	Lost	Topic 3	Topic 1	Topic 0	0.25
44	Somewhere Over The Rainbow/What A Wonderful World	Topic 0	Topic 4	Topic 4	0.38

- A discussion comparing the results of the two approaches, reflecting on their success at addressing your task or question, and identifying any shortcomings or things you might decide to try with more time. As you consider possible shortcomings, you should consider what potential sources of bias there might be in both approaches, along with what potential impacts such bias might have if employed in a real-world context.

  Results statistics for spotify data with different number of passes

  Let n represent the number of passes at each implementation of the topic model.

	n = 5	n = 10	n = 15	n = 20	n = 25	n = 30
Perplexity	-7.58	-7.56	-7.55	-7.55	-7.55	-7.52
U_Mass Coherence	-2.97	-3.39	-3.72	-3.72	-3.71	-3.38
U_uci Coherence	-0.08	-0.29	-0.62	-0.61	-0.55	-0.32
Topic diversity	0.84	0.92	0.94	0.95	0.95	0.97
Topic Consistency	-0.93	-0.99	-1.00	-1.01	-1.01	-1.02
Topic Interpretability	0.18	0.19	0.19	0.19	0.19	0.19

From the table above, we can see how the various evaluation metrics vary for an increasing number of passes. The perplexity score of the topic model decreases slightly with a bigger number of passes, suggesting that the model is getting slightly better at handling unseen data as the number of passes increases. This makes sense because a greater number of passes indicates the model has been trained on the corpus a greater number of times, hence increasing its performance. Similar trends can be seen with all the other metrics, whose values project a better performance as the number of passes increases, although the increase in the evaluation scores are very slight.

Human Evaluation Analysis for the model trained with 10 passes

Two evaluators rated the coherence of the words identified by the topic for each topic on a scale of 1-10, with 10 being the most coherent score. The ratings reflect how well the words fit together in the context of a given topic. The values of the ratings were averaged as displayed in the *avg. word coherence* column. The evaluators were then asked to assign labels to the different topics produced by our topic model, as well as rate from a scale of 1-10, how easy it was to assign a label to the various topics.

	Avg. word coherence	Avg. difficulty to assign labels	Labels 1	Labels 2
Topic 0	8.5	9.0	Ideas	Depression
Topi 1	4.5	6.0	Life goals	Mental reflection
Topic 2	6.0	6.5	Depression	Adventure
Topic 3	8.0	8.5	Crushing on a girl	Beach
Topic 4	6.5	7.5	A journey	Romance
Topic 5	8.0	9.5	Road trip	Family trip
Topic 6	9.0	8.5	Bad vibes	The hood
Topic 7	8.5	6.0	lifestyles	Family values
Topic 8	7.5	8.5	Romance	Work
Topic 9	8.5	9.5	Clubbing	Party
Topic 10	8.5	7.0	Travel	Going back home
Topic 11	4.0	5.0	Birthday	Going crazy
Topic 12	7.5	8.0	Conflicts	Drama
Topic 13	8.0	10.0	Heartbreak	Breakup
Topic 14	5.0	6.5	Dating	Savvy
Topic 15	8.0	7.5	Enjoyment	Playboy
Topic 16	8.0	7.5	Music	Choir
Topic 17	8.5	7.5	Body	Problems
Topic 18	7.0	7.0	An affair	Motherhood

Topic 19	7.5	9.5	Wealth	Hustle

Evaluation of Topic Model Performance

From the human evaluation of the model, it is evident that the model generally performs well in terms of how evaluators perceive the coherence of topics. The most coherent topic was Topic 9, while Topic 11 was identified as the least coherent. Many topics were relatively easier to assign labels to, as reflected by high scores for topics like Topic 13, Topic 19, Topic 9, and Topic 5. Conversely, Topic 11 proved to be more challenging to label, which aligns with its lower coherence score. Social and location-based topics were the easiest to label, while abstract concepts proved more difficult. This is particularly true for Topic 11, which had a high degree of abstraction and low coherence. Family and relationship themes, on the other hand, demonstrated high labeling consistency, suggesting that the model was able to capture these themes effectively and group them under the same category. It is important to note that the labels assigned by evaluators were purely subjective and based on the top words displayed for each topic. While the labels varied across topics, many were quite similar for certain topics, particularly Topic 9, the most coherent topic. This shows that the model performs decently in classifying topics into consistent categories, though the inherent subjectivity of music must be considered. However, some labels assigned by the evaluators to certain topics are quite similar, such as Topic 5 and Topic 10, which both consistently reflect a theme of a journey. This suggests that the model may not be optimally distinguishing between topics. This could be observed in the visualization, where the circles representing topics overlap, indicating the presence of common words across multiple topics.

Music interpretation is highly subjective, shaped by personal experiences, cultural background, and individual perceptions. Given this, the challenge remains: how well can a topic model classify song lyrics in a way that aligns with human interpretation? The evaluators play a crucial role in determining the model's performance. It is essential to consider how evaluators are chosen, how many are involved, and their potential biases based on personal experiences, age, background, and musical preferences. In this study, the evaluators were two college students, which limited the diversity of perspectives. Due to time constraints, we could not involve additional evaluators, leading to potential bias in the data, as it is largely shaped by the views of just two individuals. While the topic model demonstrates promise in categorizing song lyrics, its success is influenced by subjective human interpretation, and the limited evaluator sample introduces potential bias. Further studies with a more diverse group of evaluators could provide deeper insights into the model’s effectiveness in music topic classification.

Comparison of the results of the two approaches, reflecting on their success at addressing our task, and identifying any shortcomings or things we might decide to try with more time.

There was a lot of variation between the categorization of the testing dataset made by the topic model we implemented, ChatGPT, and our own categorization; which was expected. In categorizing these songs, each party had access to context the others did not. There were a couple cases in which the three classifications agreed, but a majority of them were 2 parties having the same placement and the third one being different. In terms of success, the research proved effective, with no notable failure. This research further cemented the inherent subjectivity in both categorization and music interpretation.

In terms of factors that could have led to differences in the distribution of the testing dataset of songs across the topics generated by our topic model, the obvious one is the advantage we have as humans. As mentioned before, the LDA model had nothing but the lyrics and the processed math to make these judgements. As humans on the other hand, we can reference our entire life experience to make these categorization decisions. When analyzing these songs, we consider not only the lyrics but also the tone of the artist, the emotion in their voice, and how the song connects to our own personal experiences. ChatGPT and the gensim model on the other hand, interpreted these songs at face value by the lyrics in the songs alone. It would be near impossible for ChatGPT to understand the emotions brought via the tone of the artist (at least at this point) or to separate our human biases from the songs in order to get a more even comparison. The factors that play into ChatGPT's categorizations are currently ambiguous to us. As a large language model, and not a topic model, it is hard to point out the different parameters it considered whilst it made its categorizations. Additionally, we factor in the context of how words are used, lyrical trends, slangs, and more. These elements are highly subjective, which may lead us to interpret or classify songs under different topics than the language models might, as they might not necessarily have an in-depth understanding of these nuanced aspects. However, in an effort to quantify these aspects, we utilized Spotify's API, which provided statistics for songs across categories such as danceability, liveness, and others. Those, however, showed us that these kinds of attributes remain just as nuanced when it comes to music even when a big corporate company attempts to standardize them.

	artist	album	track_name	track_id	danceability	energy	key	loudness	mode	speechiness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature
0	The Marías	Submarine	No One Noticed	3siwsiaEoU4Kuuc9WKMUy5	0.70	0.34	7	-10.61	1	0.03	0.08	0.12	0.46	97.99	236907	4
1	Rihanna	Good Girl Gone Bad	Don't Stop The Music	1Jo0Zg7XlrA6z0mFTZVdkn	0.83	0.67	6	-5.58	0	0.06	0.00	0.05	0.55	122.67	267080	4
2	Gus Dapperton	Orca	Palms	4NC020Nl4UqphmIt0PByw	0.75	0.56	4	-8.34	0	0.03	0.09	0.13	0.84	117.97	240240	4
3	MAGIC!	Don't Kill the Magic	Rude	6RtPijgfPKROxEzTHNRiDp	0.77	0.76	1	-4.99	1	0.04	0.00	0.30	0.92	144.03	224840	4
4	Tory Lanez	Alone at Prom (Deluxe)	Poison Ivy	2gdz7TlVOlDemmW2RoI3st	0.75	0.79	1	-6.26	1	0.20	0.00	0.10	0.75	106.08	187924	4
5	Ozzy Osbourne	No More Tears (Expanded Edition)	Hellraiser	4nvYDy427VvA1TcEkdgZII	0.40	0.85	4	-6.35	1	0.03	0.00	0.23	0.51	96.06	292720	4
6	Various Artists	Teen Beach Movie	Cruisin' for a Bruisin' - From "Teen Beach Movie"/Soundtrack Version	5e3UWcbvTaF5CJWu81MqlQ	0.58	0.91	11	-4.62	1	0.12	0.00	0.11	0.78	163.93	195733	4
7	Linkin Park	Hybrid Theory (Bonus Edition)	In the End	60a0Rd6pjrkxjPbaKzXjfq	0.56	0.86	3	-5.87	0	0.06	0.00	0.21	0.40	105.14	216880	4
8	Natasha Bedingfield	Unwritten	Unwritten	3U5JVgI2x4rDyHGObzJfNf	0.71	0.80	5	-6.33	1	0.04	0.00	0.08	0.63	100.01	259333	4
9	Dave	Sprinter	Sprinter	2FDTHlrBguDzQkp7PVj16Q	0.92	0.68	1	-4.70	1	0.20	0.00	0.06	0.71	139.06	229133	4
10	J. Cole	Born Sinner (Deluxe Version)	Crooked Smile (feat. TLC)	5gFoAVTN9YIlM9uJCrFZtgI	0.61	0.78	7	-6.29	1	0.27	0.00	0.69	0.48	80.98	278573	4
11	Daniel Caesar	Freudian	Get You (feat. Kali Uchis)	7zFXmv6vqI4qOt4yGf3jYZ	0.66	0.29	4	-8.53	0	0.03	0.00	0.07	0.36	74.04	278180	4
12	The Strokes	The New Abnormal	Ode To The Mets	1BLOVHYYIH4JUHQGcpt75R	0.43	0.62	1	-5.42	0	0.03	0.18	0.10	0.20	92.00	351787	4
13	JID	DiCaprio 2	Workin Out	46Lx5epWOQBA3J86ovnxmV	0.82	0.57	10	-8.07	0	0.28	0.00	0.19	0.56	126.94	226859	4
14	The Alchemist	Bread	E. Coli (feat. Earl Sweatshirt)	3GXSywNvYLAVUCtjMHkKDh	0.32	0.56	11	-8.73	0	0.13	0.00	0.32	0.40	96.02	132147	3
15	SZA	SOS	Snooze	4iZ4pt7kvcaH6Yo8UoZ4s2	0.56	0.55	5	-7.23	1	0.13	0.00	0.11	0.39	143.01	201800	4
16	Rustage	Survive (Ban Rap) [feat. DaisyBanaisy]	Survive (Ban Rap) [feat. DaisyBanaisy]	5HlMk5zD8jHLY5cA1BimjJn	0.69	0.69	9	-7.31	1	0.29	0.00	0.09	0.20	149.73	204608	4
17	Hillsong Worship	No Other Name (Deluxe Edition/Live)	Broken Vessels (Amazing Grace) - Live	2BuhGnXpqMQ8BvOUZpmaVl	0.38	0.51	7	-6.48	1	0.03	0.00	0.72	0.09	140.99	568787	4
18	Nicki Minaj	Last Time I Saw You	Last Time I Saw You	79DPYZ6x8FfzgSgPoPhpC3	0.84	0.43	5	-6.27	0	0.04	0.07	0.09	0.30	129.98	216348	4
19	Normani	Motivation	Motivation	0rlAC4PXANcKmtJfoqmVm	0.60	0.89	4	-3.97	1	0.10	0.00	0.30	0.88	170.92	193837	4
20	Fall Out Boy	American Beauty/American Psycho	Centuries	04aAxqtGp5pv12UXAg4pkq	0.39	0.86	4	-2.87	0	0.07	0.00	0.10	0.56	176.04	228360	4
21	Thousand Foot Krutch	The End Is Where We Begin	Courtesy Call	0AOmbw8AwDnwXhHC3OhdVB	0.53	0.64	11	-5.14	0	0.08	0.00	0.08	0.44	164.08	236898	4
22	Ariana Grande	My Everything (Deluxe)	Best Mistake	70yhaHLp9STtzI2Kzba6Tr	0.65	0.58	6	-6.90	1	0.45	0.00	0.11	0.23	143.87	233733	4
23	Westlife	Back Home	Home	7BbZiTZdEF7enOkY6BmUV2	0.55	0.55	10	-5.42	1	0.03	0.00	0.09	0.17	122.02	206680	4
24	Adele	25	Send My Love (To Your New Lover)	0t7fVeEJxO2XI4H2K5Svc9	0.69	0.53	6	-8.36	0	0.09	0.00	0.17	0.57	164.07	223079	4
25	League of Legends	Legends Never Die	Legends Never Die	1FpVJ7HpZlnE2GvhVE2TwT	0.50	0.60	4	-6.64	0	0.04	0.00	0.11	0.06	140.08	235000	4
26	Lil Wayne	Tha Carter IV (Complete Edition)	How To Love	5W7BcYDIaeXnpG3e39T7hJ	0.64	0.66	11	-6.09	1	0.04	0.00	0.11	0.27	153.99	240307	4
27	Ella Mai	READY	Boo'd Up	0Al2qAoaVjlwVNAqrscXli	0.56	0.78	10	-5.11	0	0.05	0.00	0.08	0.24	81.96	256064	4
28	Miley Cyrus	Bangerz (Deluxe Version)	Adore You	5AnCLGg35zIFOIoEnXK4uu	0.58	0.66	0	-5.41	1	0.03	0.00	0.11	0.20	119.76	278747	4
29	Chord Overstreet	Hold On	Hold On	5vjLSffimiIP26QG5WcN2K	0.62	0.44	2	-9.68	1	0.05	0.00	0.08	0.17	119.95	198853	4
30	Ghost	Seven Inches of Satanic Panic	Mary On A Cross	2wBnZdVWa5jVpvYRfGU7rP	0.46	0.90	11	-4.46	1	0.05	0.00	0.11	0.56	129.99	244804	4
31	Thundercat	Drunk	Them Changes	7CH99b2i1TXS5P8UUyWtnM	0.66	0.56	8	-8.60	1	0.06	0.00	0.10	0.70	81.66	188454	4
32	Tory Lanez	Alone At Prom	The Color Violet	3azJifCSqg9fRIj2yKIbWz	0.64	0.53	6	-10.80	0	0.05	0.00	0.09	0.46	105.02	226467	4
33	Nickelback	All the Right Reasons (Special Edition)	Rockstar	6n9yCXvLhnYMgJIiIcMu7D	0.62	0.91	0	-3.00	1	0.04	0.00	0.34	0.69	144.07	252040	4
34	Nitty Gritty Dirt Band	Hold On	Fishin' in the Dark	15jhnquhNS1IZMlM493UBW	0.74	0.35	2	-13.14	1	0.04	0.00	0.09	0.90	155.54	201600	4
35	Dominic Fike	Don't Forget About Me, Demos	3 Nights	1INJrcVe6gwLEIZCtprs1u	0.82	0.52	7	-6.59	0	0.09	0.00	0.10	0.88	151.89	177667	4
36	Djo	DECIDE	End of Beginning	3qhIB30KknSejmIvZZLjOD	0.69	0.45	2	-7.64	1	0.06	0.00	0.07	0.91	159.98	159246	4
37	Eagles	Legacy	Hotel California - 2013 Remaster	4yS9mjJiE8dyC8EVg9vupQ	0.58	0.51	2	-9.48	1	0.03	0.00	0.06	0.61	147.12	391376	4
38	Glen Campbell	Southern Nights	Southern Nights	7kv7zBjMtVf0eIJle2VZxn	0.70	0.81	11	-8.89	0	0.03	0.02	0.58	0.85	95.30	180027	4
39	Billie Eilish	HIT ME HARD AND SOFT	BIRDS OF A FEATHER	6dOtVTDdiauQNBQEDOtlAB	0.75	0.51	2	-10.17	1	0.04	0.06	0.12	0.44	104.98	210373	4
40	Wild Cherry	Wild Cherry	Play That Funky Music	5uuJruktM9fMdN9Va0DUMI	0.81	0.67	9	-12.07	1	0.06	0.00	0.06	0.93	109.39	300000	4
41	Sade	The Best of Sade	Smooth Operator - Single Version	1Hv1VTm8zeOeybub15mA2R	0.73	0.58	9	-6.62	0	0.03	0.00	0.03	0.96	119.34	258693	4
42	Creed	Human Clay	Higher	1ZozJfi8u9cO2Ob8KwiwNT	0.46	0.83	2	-6.25	1	0.04	0.00	0.21	0.43	155.83	316733	4
43	Frank Ocean	channel ORANGE	Lost	3GZD6HmiNUhxXYf8Gch723	0.91	0.60	8	-4.89	1	0.23	0.00	0.17	0.50	123.06	234093	4
44	Israel Kamakawiwo'ole	Facing Future	Somewhere Over The Rainbow_What A Wonderful World	25U7raB3ZSszayTYCIh4hF	0.66	0.17	0	-13.72	1	0.04	0.00	0.36	0.68	85.00	308027	4

Fig: It is interesting to see how the features of our test songs compare to the average for the topic it was placed into. In some ways they're one in the same and in others they're outliers.

Despite the general success of our research, there are many other aspects that would have been worth exploring given more time and technical resources. For instance, the possibility of training a topic model to include more inherent criterions such as danceability, loudness, mode, and so on. Finding ways to include these factors in topic generation, as well as song categorization sounds like a great task for future exploration. On another note, our model was trained with song lyrics, but that's not all humans are exposed to. We are also curious as to how different this model and these results might have turned out had the model been trained on modern day conversations alone, or both.

Given the variety of systems Spotify employs for song recommendations, we aimed to explore how effectively natural language processing (NLP) algorithms could categorize songs based on various criteria, including but not limited to song lyrics. After conducting our research and comparing the results of topic modeling against classifications made by ChatGPT and human evaluators, we concluded that while both topic modeling and LLMs like ChatGPT perform reasonably well in classifying songs, a significant degree of subjectivity remains in determining how well songs fit into the assigned categories. Furthermore, lyrics alone may not fully capture the emotions conveyed by a song, which can heavily influence classification decisions. While our implementation does a decent job in categorizing songs from the testing data into distinct topics, there is room for improvement and optimizing topic consistency, diversity, and interpretability. These enhancements could help reduce overlaps between topics

and result in more precise classifications. Overall, while our approach demonstrates potential, additional work is needed to refine the model for better real-world usability.