Sentiment Analysis of Amazon Reviews

1. Data description

- **How and when the data were collected:**

  The data we use are CDs and Vinyl Review data and Metadata of the Amazon Review Data in 2018. The Amazon Review Data in 2018 an updated version of the Amazon Review Dataset in 2014. The data were gathered by Jianmo Ni and Julian McAuley's lab at UCSD. The specific way that the data were collected is unknown to us, probably sponsored by Amazon.

- **What are the shortcomings of the data (e.g., possible sample biases), if any?**

  This dataset includes all reviews for CD and Vinyls department in 2018, which means some niche albums with only one review were also counted. In addition, the dataset lacks customer demographics, making follow-up personalized recommendations harder.

- **Data size and dimensions:**

  The dataset includes 2 main parts for each category: reviews and product metadata. There are complete data including all categories at Amazon, as well as per-category data which includes only the review and product metadata for each category. We only choose CDs and Vinyl Review data and Metadata to perform our analysis.
  In the review data, there are 4,543,369 reviews (observations) in total, and for each reviewer, there are 12 variables recording the information such as the name of the reviewer, the votes of the views, the rating of the product, etc. The review data don't include information on the product, but record IDs of the products that can be linked to the product metadata. There are 17 related product details in Product Metadata including descriptions, category information, price, brand, image features, etc.

2. Project objectives

- Determine and visualize main topics in the CDs and Vinyl reviews.
- Evaluate models with different topic numbers and examine the best topic model.
- Categorize the sentiment of the CDs and Vinyl reviews.
- Find out what types of audience are more likely to give higher scores and more positive reviews.
- Extract aspects from reviews to know what people care about when they are buying CDs and Vinyl.
- Score the reviews based on aspects extracted to see how people evaluate the product.

3. Methodology

We split the problem into 2 main parts: topic modeling and sentiment analysis. We choose Latent Dirichlet Allocation as our first technique to extract topics. For sentiment analysis, we first compare 2 methods (Support Vector Machine and Lexicon) of sentiment analysis, and then use the VADER scores we get from sentiment analysis to run a regression to find out which topic people

who are more likely to give high scores and comments overall pay more attention to. Lastly, we try another sentiment analysis method based on Bing Liu's Opinion Lexicon to extract aspects, score reviews and generate more insights.

Data Cleaning

After reading the two json files, we remove rows with null values and unstructured useless information (i.e. unformatted title may contain html style content).

We merge two pre-processed datasets CDs and Vinyl Review data and Metadata, and then select related columns like *"ReviewText", "vote", "verified", "price", "brand", "asin"* to further preprocessing. To increase the credibility of our analysis, we focus on the dataset with more than one comment and verified purchase.

Lastly, we extract the *"ReviewText"* and normalize it by converting all words to lowercase and removing punctuations and HTML tags and stop words.

Topic Modeling

Before working out with topic modeling, we first need text data in Bag-of-Words representation. So, we vectorize our normalized corpus and limit the number of features to 1000 most frequent features for further steps.

Without knowing what topics could appear in more than 4 million reviews, we randomly select 100,000 reviews and run the Latent Dirichlet Allocation function for 3 topics first. We set the maximum iteration to 100 for faster calculation. After classifying words into topics based on relevance, we calculate each word's weight and generate a word vector for every topic. We can visualize each topic's size and frequency of words with the function pyLDAvis after sorting word weights. In order to find the best number of topics, we fit the corpus with 5 and 7 topics again with repeating steps we do for the 3-topic model. The next step is to determine each review's dominant topic. By summing weights for a particular topic across all words in a review, we can get the weight of that topic in the review.

Having the Bag-of-Words formatted corpus with each review and top 20 words for each topic in hand, we can calculate the coherence by the overall model and by each topic. We can get log-likelihood scores and perplexity scores as well to further evaluate which topic model has the best classification.

Finally, by adjusting the value of lambda (0, 0.5, and 1) for the best topic model, we conclude the name for each topic based on the top-30 most relevant words with the selected lambda.

2 Methods of Sentiment Analysis

Method 1:

We choose 1 supervised modeling and 1 unsupervised modeling to compare the accuracy of the models. The supervised model we choose is Support Vector Machine and the unsupervised is VADER Lexicon through NLTK module. In the original dataset, we have a total score for each comment from 1 to 5 stars.

In this case, binary polarity is not suitable, so we use “positive”, “neutral”, and “negative” as human-created labels to recategorize the score column.

After randomly splitting the dataset into training and test, we extract features, use the features as dependent variables and the polarity we get from the scores as the outcome variable to train an SVM Classifier on training data, and then predict the polarity of test data.

Using the unsupervised classification model Lexicon, we first fit the model on training data and assign binary sentiment polarity with a threshold of 0.1. Then we use human-created labels and compute the crosstab made by VADER Lexicon on test data with the threshold value of 0.1, which turns out to be quite accurate. We adjust the threshold parameter to fine-tune the accuracy rate, precision and recall rate to get the maximized accuracy rate, with three plots showing that most comments are positive.

Fixed-Effect Regression on Topics

After topic modeling and sentiment analysis, we now get the weights of each topic and also the polarity score of each comment. We run a regression on compound scores from sentiment analysis using the weights of 5 topics including “format and quality”, “classic music”, “appreciation”, “genre” and “general evaluation”, while controlling the fixed-effect regression within each artist, because different artist may have different genre and different audience, which makes the topic incomparable.

Method 2: Sentiment analysis based on aspects

a. Extract Aspect
   We want to focus on what aspect each review mentions. Here, we extract aspects by filtering out 100 nouns that appear most frequently. Then we manually select 60 of them based on common sense. For example, “year”, “release”, “something” are not the aspects we want to focus on. We want to analyze the “quality”, “performance”, “sound” of the CD.
b. Read Lexicons
   Here we use Bing Liu's Opinion Lexicon to score sentences. There are positive lexicons and negative lexicons.
   (source: Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA)
c. Predict Score
   We try to measure sentiment score by the proportion of absolute number of positive words(number of positive words minus number of negative words) to number of words in tokenized sentences.
d. Get Score by Aspect
   For each aspect, for example “rock”, we can extract all reviews that mentioned this aspect and score all the reviews to know how people think of the aspect. For instance, how do people care about “rock” music compared to “jazz”.
e. Analyze review by product
   We can filter out all reviews of one product and analyze it’s score on specific aspects.

4. Results

Results of Topic Modeling

After comparing coherence scores, log-likelihood scores, and perplexity scores for 3, 5, and 7 topic modeling respectively, we choose 5-topic modeling because it performs relatively better than the other two topic models. (5-topic modeling has the best coherence score and perplexity score. Its coherence score also ranks second among 3 models. All measures of performance are shown in the appendix)

Then we go back to the 5-topic visualization to adjust the lambda. We find that as lambda increases, more general words without meaning are included in each topic. But if the lambda is small, top words are too exclusive to the topic they belong to. Thus, when lambda equals to 0.5, top words in each topic are semantically related to each other the best. We conclude the name for each topic based on its top-30 most relevant words as follows (word order follows the estimated term frequency within the selected topic):

Topic 1: General Evaluation (25.3% of tokens)

album, like, song, really, good, think, lyric, feel, thing, something, sound, track, first, much, metal, way, even, beat, start, bad, still, lot, seem, little, bit, maybe, pretty, different, kind, heavy

Topic 2: Classic Music (20.7% of tokens)

recording, performance, work, piece, piano, music, orchestra, symphony, musical, concerto, instrument, composer, movement, score, opera, string, major, classical, de, theme, listener, composition, la, play, perform, violin, minor, sonata, fine, interpretation

Topic 3: Genre (19.3% of tokens)

band, rock, guitar, blue, jazz, tune, hit, classic, group, vocal, live, great, solo, drum, record, top, bass, show, best, john, early, roll, musician, year, lead, feature, late, man, 70s, king

Topic 4: Appreciation (18% of tokens)

love, music, cd, song, voice, great, sing, enjoy, buy, hear, favorite, beautiful, wonderful, every, year, recommend, old, Christmas, artist, time, amazing, friend, day, know, anyone, thank, car, heart, singer, life

Topic 5: Format and Quality (16.8% of tokens)

version, original, set, dvd, disc, cd, quality, sound, release, good, vinyl, movie, price, amazon, video, worth, collection, lp, concert, box, track, soundtrack, copy, include, bonus, available, record, stereo, audio, live

Results of Sentiment Analysis

Method 1:

The accuracy rate using supervised model SVM is around 0.864. The maximized accuracy rate using unsupervised VADER Lexicon-based model is around 0.85. We use -0.415 as our final threshold value which yields the maximum accuracy rate to get the final predicted scores of each comment in the test data. In this case, we find that supervised models may have a few more advantages than unsupervised models on predicting polarity when there is already an outcome variable such as overall score, but the lift is not much.

Results of Fixed-Effect Regression

The coefficients of “classic music” and “appreciation” are the highest. It shows that people who pay more attention to “classic music” and “appreciation” are more likely to give higher scores to CDs or vinyl products.

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	score
No. Observations:	88177	Method:	REML
No. Groups:	22562	Scale:	0.1167
Min. group size:	1	Likelihood:	-32960.0056
Max. group size:	2817	Converged:	No
Mean group size:	3.9		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.538	21529.928	-0.000	1.000	-42199.420	42196.345
quality	2.048	21529.928	0.000	1.000	-42195.835	42199.931
classic	2.473	21529.928	0.000	1.000	-42195.410	42200.356
appreciation	2.483	21529.928	0.000	1.000	-42195.399	42200.366
genre	2.343	21529.928	0.000	1.000	-42195.540	42200.226
eval	2.220	21529.928	0.000	1.000	-42195.663	42200.103
Group Var	0.011	0.002				

Method 2:

Exhibit 1: How do people say about “song/lyrics” in all reviews?
Compared to people who mentioned "song" in their reviews, people mentioned "lyric" are very few. The reviews mentioned “song” are less positive than reviews mentioned “lyrics”, which might indicate that people love the song itself better than the lyrics.

Exhibit 2: How do people say about “rock/jazz” in all reviews?
Comparing to people who mentioned "rock" in their reviews, people mentioned "jazz" are very few. Meanwhile, people's preference towards rock is a little bit lower than jazz. This tells us that rock is more popular among people but people's overall preference to jazz is higher than rock music.

Exhibit 3: How do people say about “quality” in all reviews?
We used to think when people mentioned quality, they must be complaining about it. From the distribution, we can see it's not the case. The majority of people mentioning quality provide positive review (score higher than 0). People like to praise the quality of a CD rather than just complaining about it.

Exhibit 4: How do people say about the CD *I Dreamed A Dream* in all reviews?
I Dreamed A Dream is the CD that has the most reviews in our dataset. It was released on 23 November 2009. In only six weeks of sales, it became the biggest selling album in the world for 2009. From our scores, we can also tell that people like it. Lots of people buy it and most of them hold a positive view on it.

5. Conclusions

Recommendations Based on the Results of Topic Modeling

Classic music takes up a large share in topic modeling, which shows classic music is a huge market. Compared with today's on-demand music streaming world, classical music audiences are more likely to purchase physical albums in the form of CDs and Vinyl. So we suggest Amazon grasp this classical music resurgence to stock theses classic albums and introduce technology that serves these audiences better.

Through the exploratory analysis of Topic Modeling Format and Quality, we find that the record format requirements of the general music audience are various, such as CD, LP, DVD, etc. And they also have a higher demand for good quality of records, especially for soundtracks. Thus, the urgent action for Amazon is to put more high-resolution audiophile-grade recordings into its CD and Vinyl department and provides diverse formats capable of storing high-resolution audio.

As words for physical albums (like CD and LP) show high relevancy and frequency in topic 5, CD players and vinyl record players also have good prospects of sales. We recommend Amazon can show cd players and vinyl record players in the section of " customers who bought this item also bought" when a customer is searching for physical albums. This improvement can also promote the sales of physical album's players.

Topic Modeling analysis reveals that almost all reviews are positive and neutral evaluation and seldom bad evaluations show, which implies that Amazon does a super great job at the CD and Vinyl market. On the one hand, it may be because CD and Vinyl is a niche market, only true music lovers will devote money to it. Their satisfaction and appreciation strengthen our conclusion that Amazon is phenomenal. On the other hand, we need to consider the possibility of positive bias in reviews. There may exist the possibility that dissatisfied customers just churn to other platforms and reject to express their negative true evaluation.

Shortcomings in Topic Modeling and Possible Ways to Overcome

There are polysemous words and same words in different contexts in our review. When we interpret these words out of context, their exact meanings are hard to get. One way to overcome it is to take the whole context topic into consideration by using LDA.

Due to the large CD review dataset, we randomly sample 100000 reviews to conduct LDA analysis, which may make our model less comprehensive. But we rerun the sampling process several times and utilize different datasets to detect the proper topic model number. And our results are super similar, which means our conclusion is plausible. But if possible, we still prefer using all the reviews data to do the topic modeling.

Parameter selections are subjective. The best topic number and max interaction number depend on multiple experiments. Topics are soft-clusters so there is no objective metric to say "this is the best choice" of hyperparameters. And different evaluation rules emphasize different aspects. So we need to make a trade-off when they perform well in some areas but poorly in other areas. Metrics like perplexity (how well the model explains the data) are okay to test if the learning is working, but very poor indicators of the overall quality of the model. Also, it is subjective to generalize what the topic is talking about with very clear semantics. A better way to improve is that we can use some unsupervised clustering method to divide the text into some clusters, establish the corresponding relationship between these clusters and category system, then review these clusters manually, segment clusters and then build a hierarchical taxonomy according to the knowledge system.

Recommendations Based on Sentiment Analysis

Based on the results of the regression, companies could use the insights to determine which kind of customers they should focus on obtaining, for example, they could try to attract more people who love classical music and continue to develop marketing strategies for them.

Based on the results of the second sentiment analysis method, we came up with these recommendations: First, when comparing products, companies can select some specific aspects to compare products to generate a whole picture of how the products are competing with or different from each other. Second, companies can know customers' attitude towards each aspect, thus improving the product on that aspect if customers are not satisfied with it. Third, for certain customers, companies can also extract their reviews to see what aspects they are paying most attention to and conduct targeted marketing based on their personal preference.

Shortcomings in Sentiment Analysis and Possible Ways to Overcome

Rather than just consider the positive/negative words, when doing sentiment analysis, we should also consider how much they like/dislike, the words can show the extent they love the product like "really", "very much", "a lot" and so on.
There should be other ways to calculate sentiment score, rather than the fraction of sentiment words in the whole sentence. The approach we use is elementary.
Scores should be scaled thus we can know whether a score is high or low. For example, some reviews of *I Dreamed A Dream* are higher than 0.4, but we don't know whether it is high enough. It would be better if we have the upper limit and lower limit for scores.

6. Appendices

Topic Modeling

3 Topics Modeling Evaluation Score:

```
cm = CoherenceModel(topics=topic_topwords,
                    corpus = news_corpus_bow ,
                    dictionary = news_dictionary, coherence='u_mass')
print("Coherence score for the model: ", np.round(cm.get_coherence(), 4))  # get coherence value
```

```
Coherence score for the model:  -1.5273
```

```
print("Coherence score by topic (higher values are better): ", np.round(cm.get_coherence_per_topic(),4))
```

```
Coherence score by topic (higher values are better):  [-1.6381 -1.332  -1.6118]
```

Log-Likelihood Score

```
print("Log-Likelihood (higher values are better): ", lda_corpus_3.score(bow_data))
```

```
Log-Likelihood (higher values are better):  -26627099.232310824
```

Perplexity Score

```
print("Perplexity (lower values are better): ", lda_corpus_3.perplexity(bow_data))
```

```
Perplexity (lower values are better):  514.2328834467227
```

5 Topics Modeling Evaluation Score:

```
cm = CoherenceModel(topics=topic_topwords,
                    corpus = news_corpus_bow ,
                    dictionary = news_dictionary, coherence='u_mass')
print("Coherence score for the model: ", np.round(cm.get_coherence(), 4))  # get coherence value
```

```
Coherence score for the model:  -1.6868
```

```
print("Coherence score by topic (higher values are better): ", np.round(cm.get_coherence_per_topic(),4))
```

```
Coherence score by topic (higher values are better):  [-1.691  -1.483  -1.7701 -1.7831 -1.7069]
```

Log-Likelihood Score

```
print("Log-Likelihood (higher values are better): ", lda_corpus_5.score(bow_data))
```

```
Log-Likelihood (higher values are better):  -26558455.66116012
```

Perplexity Score

```
print("Perplexity (lower values are better): ", lda_corpus_5.perplexity(bow_data))
```

```
Perplexity (lower values are better):  506.0233830751827
```

7 Topics Modeling Evaluation Score:

```
cm = CoherenceModel(topics=topic_topwords,
                    corpus = news_corpus_bow ,
                    dictionary = news_dictionary, coherence='u_mass')
print("Coherence score for the model: ", np.round(cm.get_coherence(), 4))  # get coherence value
```

```
Coherence score for the model:  -1.9046
```

```
print("Coherence score by topic (higher values are better): ", np.round(cm.get_coherence_per_topic(),4))
```

```
Coherence score by topic (higher values are better):  [-1.9224 -2.419  -1.7435 -1.9736 -1.6618 -1.9911 -1.6212]
```

Log-Likelihood Score

```
print("Log-Likelihood (higher values are better): ", lda_corpus_7.score(bow_data))
```

```
Log-Likelihood (higher values are better):  -26583347.094135918
```

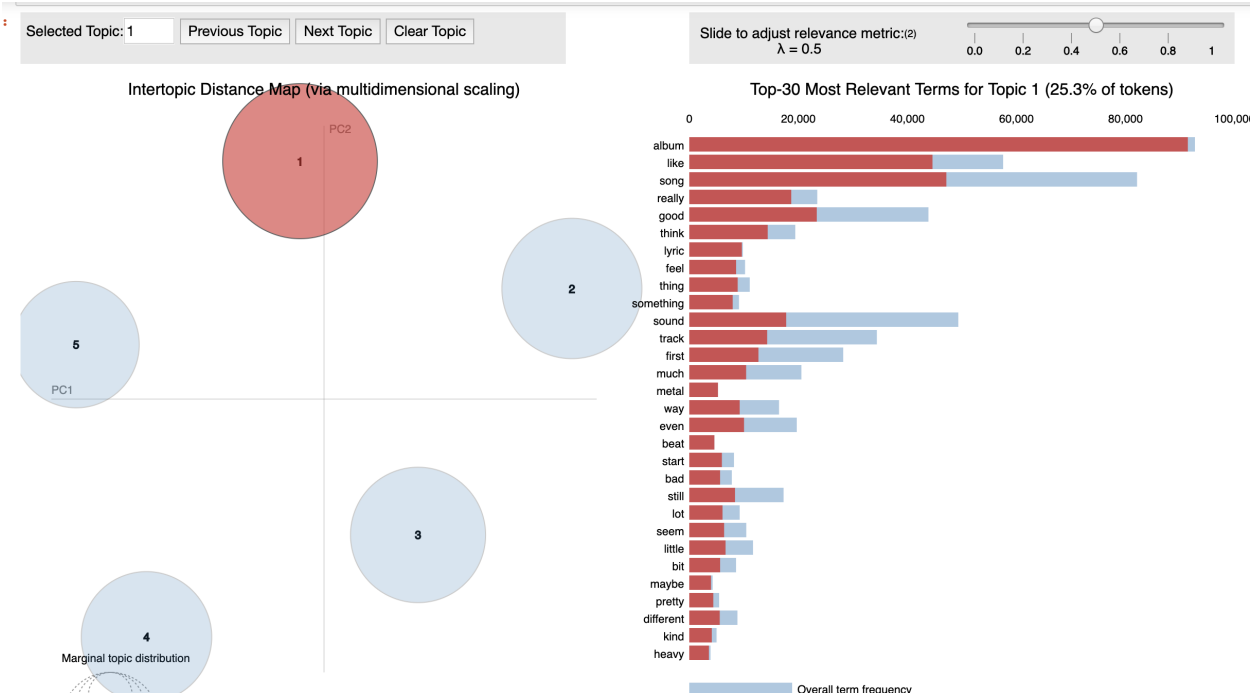
Perplexity Score

```
print("Perplexity (lower values are better): ", lda_corpus_7.perplexity(bow_data))
```

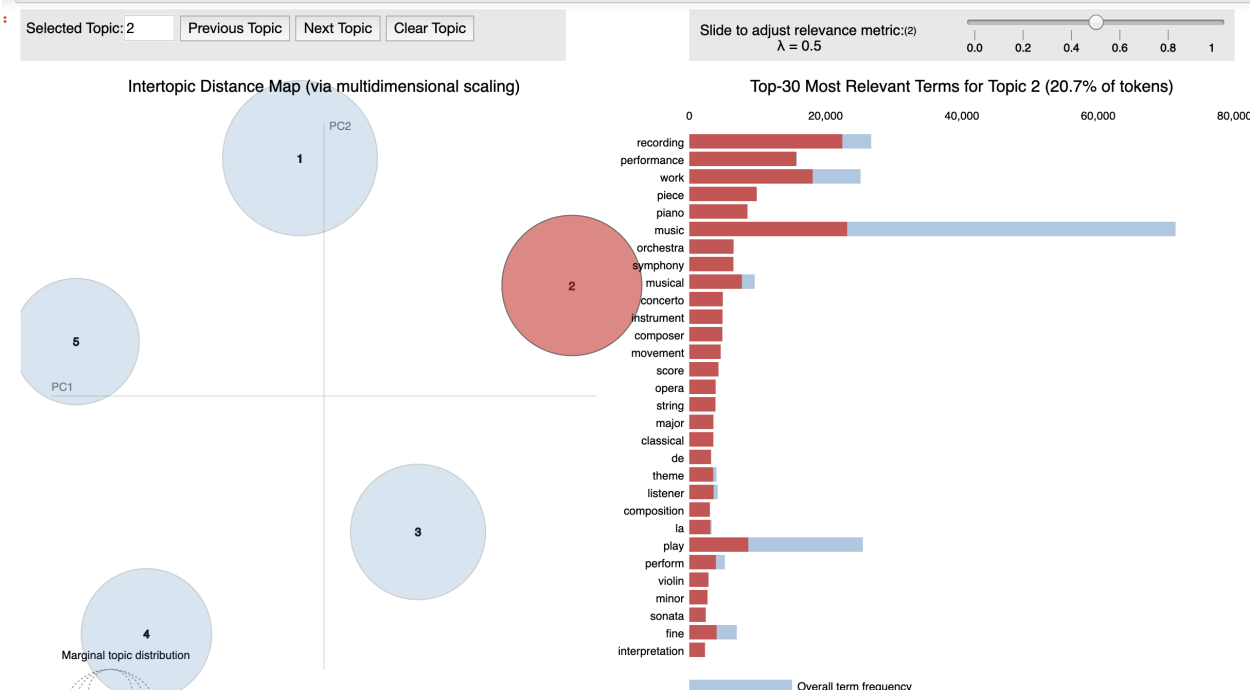
```
Perplexity (lower values are better):  508.9850435829123
```

5 Topic Modeling Visualization:

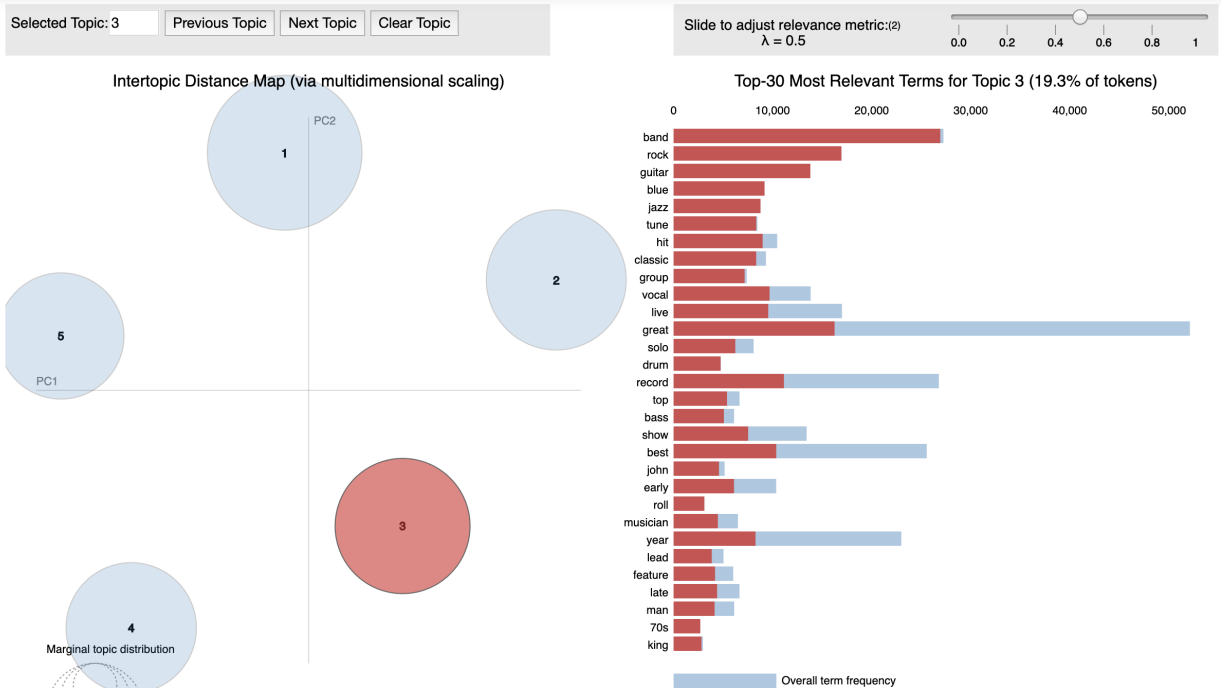
Topic 1: General evaluation



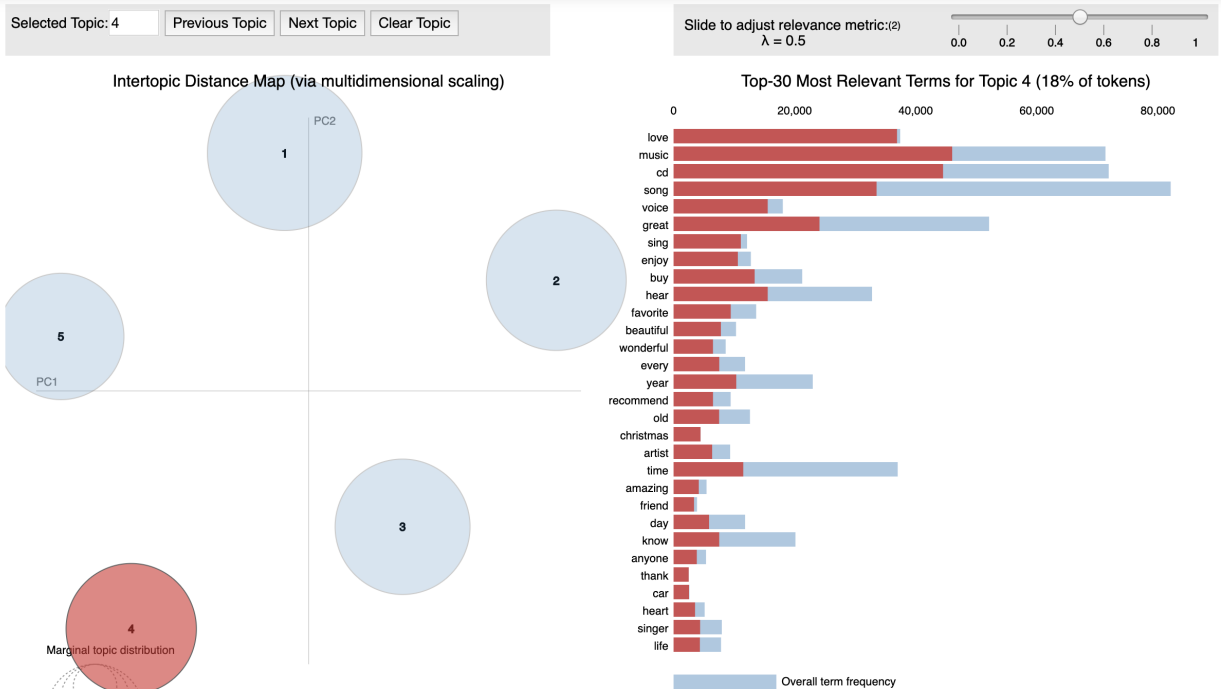
Topic 2: Classic music



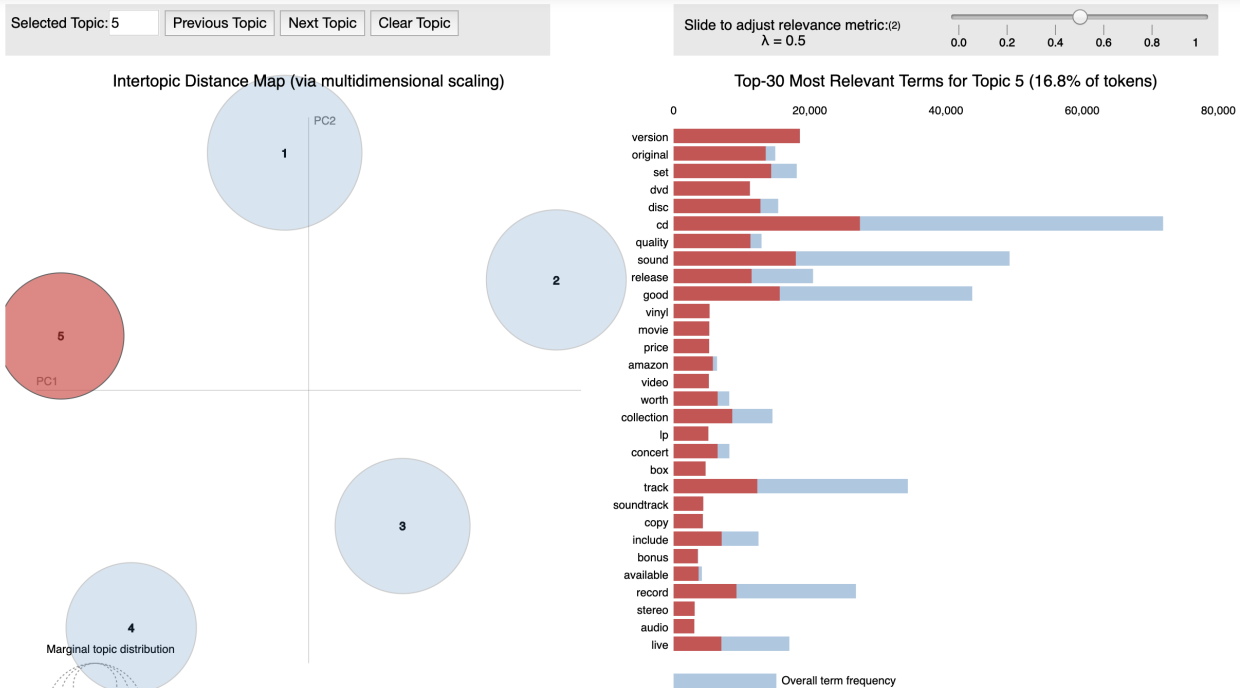
Topic 3: Genre



Topic 4: Appreciation



Topic 5: Format and quality



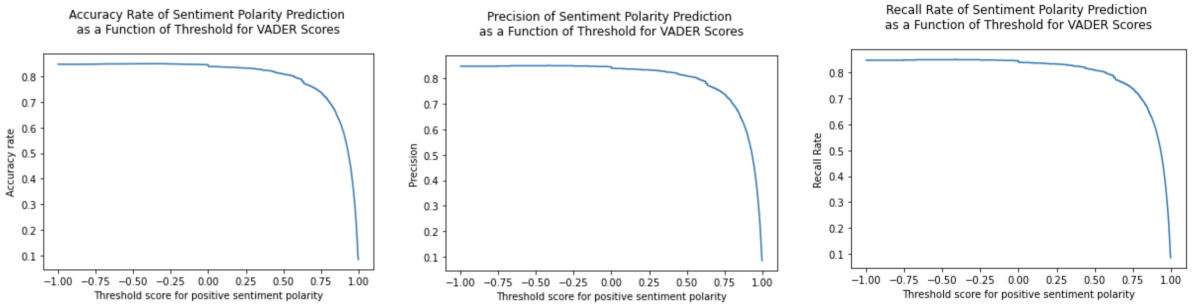
Sentiment Analysis

Method 1:

Crosstab

Predicted:	negative	positive	All
True:			
negative	90	1308	1398
neutral	79	1066	1145
positive	1025	14139	15164
All	1194	16513	17707

Plots of Accuracy Rates, Precision, and Recall Rate



Method 2:

Exhibit 1

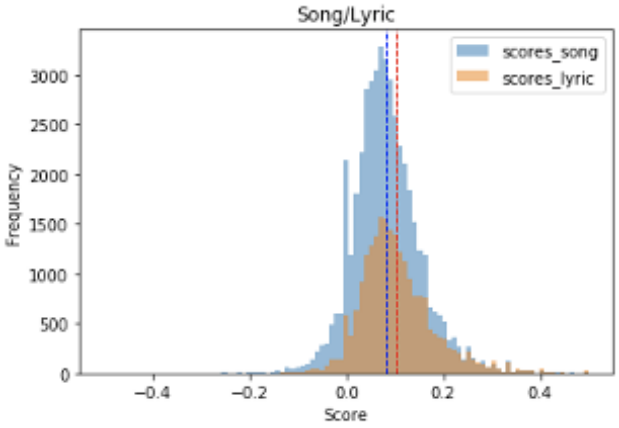


Exhibit 2

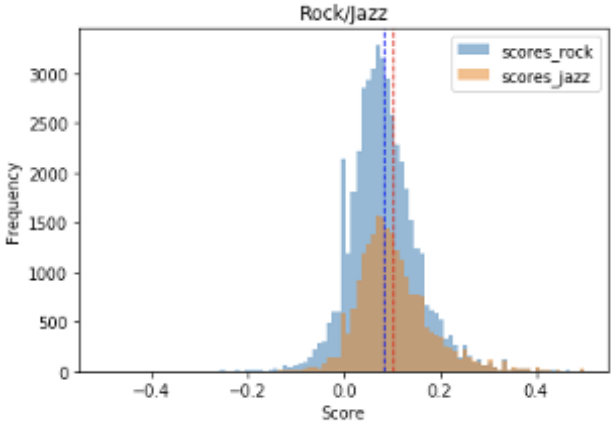


Exhibit 3

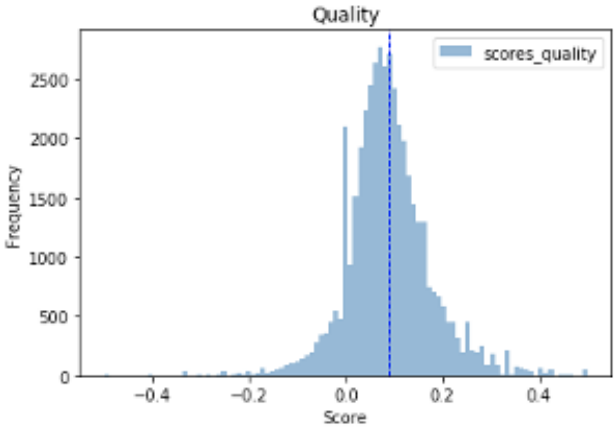
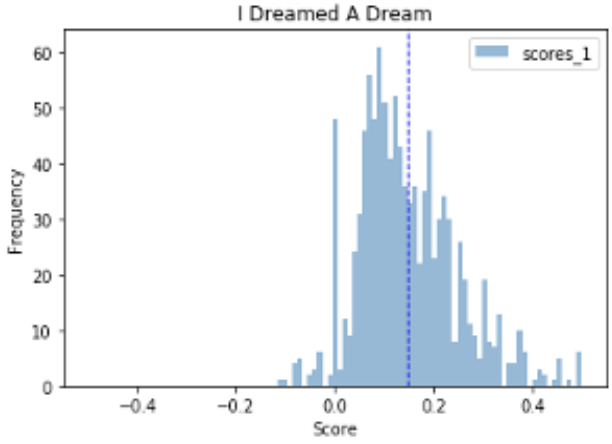


Exhibit 4


I Dreamed A Dream
scores_1
Frequency
Score
0
10
20
30
40
50
60
−0.4
−0.2
0.0
0.2
0.4