Kaur or Kendrick: An Analysis of the Overlap Between Rap and Modern Poetry Using Machine Learning

Kyle Barnes
Department of Computer Science
Princeton University
ktbarnes@princeton.edu

Saahil Katyal
Department of Economics
Princeton University
skatyal@princeton.edu

Roopa Ramanujam
Department of Computer Science
Princeton University
roopar@princeton.edu

Abstract

The divide between rap and poetry is not as clear as is commonly assumed, and our objective was to explore this relationship with supervised and unsupervised learning methods. We combined a rap dataset from Metrolyrics with approximately 34,000 raps and a modern poetry dataset from poetryfoundation.org with approximately 16,000 poems, and performed data cleaning on these as well as labeled each sample with the appropriate binary label. We then represented the data using tf-idf and performed Latent Dirichlet Allocation with two topics, as well as K-Means clustering on the resulting topic proportions. We also performed supervised learning in the form of different types of classifiers, attempting to predict both the cluster labels and the actual labels of the samples. We found that LDA outputted two distinct topics that corresponded well to rap and poetry. K-Means clustering showed that there exists noticeable overlap between rap and poetry, in that there were are raps that contain a high proportion of poetry topics than vice versa. Lastly, we discovered that logistic regression on the LDA topic feature space performed the best in terms of classifiers, reinforcing LDA as tool for dimensionality reduction. Extensions of this project include delving into the work of specific rap artists and sub-genres to further investigate these distinctions.

1 Introduction

There exists an interesting relationship between poetry and rap. To many, they seem like completely different art forms, as scholars call attention to technical differences between the two by highlighting the fact that rap is a musical and verbal art whereas poetry is musical, verbal, and typographical [1]. However, in many cases, the differences seem less clear. Rap enthusiasts point to Eminem, Jay-Z, Kendrick Lamar, and others as examples of rappers whose iconic lyrics walk the line between the two and, moreover, highlight the fact that both what we traditionally think of as "poetry" and "rap" have a word flow and aim to tell a story that serves as a reflection of the world around the artist [2]. In fact, in 2018 Kendrick Lamar was the first rapper in history to win a Pulitzer Prize, further underscoring the ambiguity surrounding rap and reviving the debate as to whether rap deserves institutional recognition traditionally reserved for art forms such as poetry, jazz, and classical music [3].

In this analysis, we leverage machine learning to address specifically the ambiguity surrounding the extent to which rap deserves to be in the same conversation as poetry. We use a data set comprised of 15,652 contemporary poems and 33,965 rap lyrics and apply Latent Dirichlet Allocation (LDA) to a term frequency-inverse document frequency matrix in order to find latent topics in the combined data set. We explore the extent to which the two classes – "rap" and "poetry" – are separated in topic space and apply k-means clustering for further analysis. Using the resulting clusters, we highlight data points whose cluster assignments differ from their true labels and seek to understand the reasons for this mismatch. This reveals underlying patterns of contemporary poems that resemble rap and vice versa. Using two topics in LDA and two clusters yielded the best results in terms of evaluation metrics and interpretability. We found that shorter, rhyming poems with heavy repetition were more commonly clustered with mostly rap lyrics, and more lyrical raps that used fewer slang terms often were clustered with poems. Lastly, we used the results from LDA to conduct classification.

This report proceeds as follows: In section 2, we discuss related work to better frame our analysis. In section 3, we describe in detail the data and methods applied. In section 4, we present our results and analysis. In section 5, we conclude by discussing shortcomings of our analysis and areas for future research.

2 Related Work

While there exists limited research that tackles our specific problem, there are a number of machine learning papers on text and natural language processing worth highlighting, as they, in part, motivate the approach to our analysis. In their analysis of topic models, Canini, Shi, and Griffiths (2009) highlight how LDA is a widely used method for identifying the topics in a set of documents given its suitability for bag-of-words data [4]. In addition, Rhody (2012) applies LDA to a body of poetry specifically and concludes that the resulting topics, despite being "opaque" with regards to interpretability, offer insight into the different ways to approach close reading of poems [5]. By revealing hidden topics, LDA allows scholars to better understand how seemingly unrelated poems draw from similar discourses [5]. We build off these papers by applying LDA to the combined data set of rap lyrics and contemporary poems to find latent topics and see how well separated the two classes – "rap" and "poetry" – are separated in topic space.

Lastly, Guan et al. (2016) highlight how k-means document clustering can be applied to LDA latent topics to improve clustering in contrast to just using k-means in the original space [6]. This motivates our choice to use k-means to cluster instances in LDA-transformed space in order to further understand the differences and overlap between rap and contemporary poetry.

3 Methods

3.1 Data Overview

In this project, we used two data sets: a data set of rap lyrics and a data set of contemporary poetry. We downloaded both data sets from Kaggle. The rap data set comes from a larger data set containing information on over 380,000 songs from MetroLyrics, a popular online lyrics database [7]. We removed all of the songs that were not designated as "hip-hop," resulting in roughly 35,000 rap songs for our rap lyrics data set.

The poetry data set originated from a larger data set containing Modern and Renaissance poetry from the Poetry Foundation [8]. We removed all poems that were designated as "Renaissance," resulting in roughly 16,000 poems in our poetry data set. We chose to focus on Modern poetry, as the word usage of Modern poets compared to Renaissance was more closely aligned with rap lyrics and, as a result, would lead to more interesting analysis.

3.2 Data Cleaning and Exploratory Analysis

We then cleaned the data sets by removing oddities in the data, such as HTML tags and markers in rap lyrics such as "Verse 1." Entries that had empty or NA Content fields were also dropped. It is important to note that the data sets included poems and raps in foreign languages, since we were not able to find a consistent way of removing non-English data. After this cleaning step, the poetry data seemed to be in better shape than the rap data, as the rap data still had some one-off oddities that we were not able to find a good way of removing. This, however, was not a big issue and did not compromise our core analysis. Our resulting data sets contained 15,652 poems and 33,965 raps, respectively.

We conducted some exploratory data analysis to gain a better sense of the data. Figures 1 and 2 show word clouds for the two data sets, highlighting specifically the most frequent words in each – excluding stop words. Figure 1 shows that words such as “one,” “love,” “light,” and “time” were the most frequent in the poetry data set. This aligned with our expectations, as they are commonly used in figurative language. Figure 2 shows the most frequent words in the rap data set, highlighting how slang words, such as “wan” and “na,” and profanity make up a significant part of rap lyrics.

Figure 1: Poetry Data Set Word Cloud

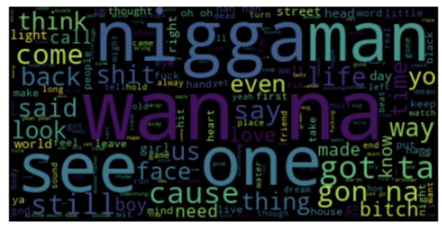

Figure 2: Rap Data Set Word Cloud

Figures 3 and 4 display histograms with the various lengths – as measured by number of words – of the poems and raps, respectively. We can see that the raps are longer than the poems in general, which was in line with our expectations.

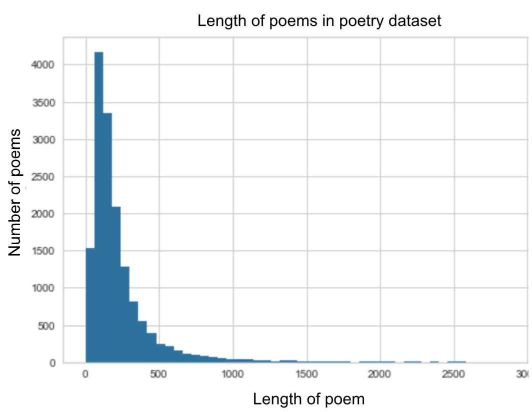

Figure 3: Length of Poems

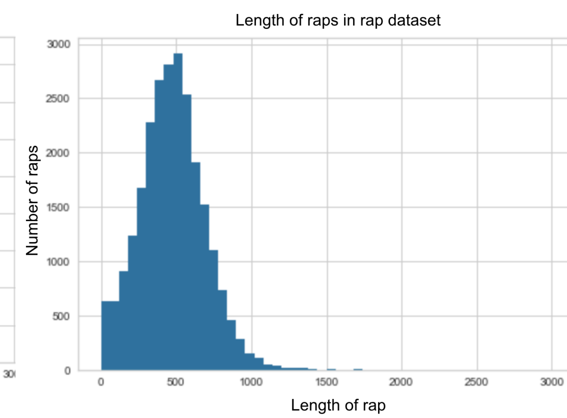

Figure 4: Length of Raps

3.3 Data Merging and Processing

We then merged two data sets by finding the common fields for both and constructing a new data set with the fields Title, Author, Content, and Label (where 0 represented poetry and 1 rap). We used two approaches to process the data: bag-of-words and term frequency-inverse document frequency (tf-idf).

We processed the poems and rap lyrics using a bag-of-words by implementing CountVectorizer from SciKitLearn [9]. More specifically, we built the vocabulary by removing English stop words and setting the maximum number of features to 1,000. This yielded a vocabulary that contained words (all converted to lowercase) we felt were both salient and interpretable.

Bag-of-words simply takes into account the frequency of words in a set of documents, and as a result, words with greater frequency are designated as more important [10]. In contrast, tf-idf serves to highlight more distinctive words in a data set [10]. As the name term frequency-inverse document frequency suggests, the vectorizer takes into account two frequencies: term frequency and inverse document frequency [10]. For a word to have a high tf-idf in a specific document, it must have a high frequency in said document, and it should not appear often in other documents in the data [10]. In short, the tf-idf is the product of these two frequencies and highlights signature words of the data set [10]. Therefore, we also processed the poems and rap lyrics using tf-idf as implemented by SciKitLearn [9]. Along the same lines described in the previous paragraph, we built the vocabulary by removing English stop words and setting the maximum number of features to 1,000 – the same size as the bag-of-words matrix. We included both uni-grams and bi-grams to help account for the fact that phrases are potentially important in song lyrics and poems. This yielded bi-grams in the vocabulary such as "just wanna," "hey hey," and "look like." At a high level, the differences between the two representations of the data seemed minimal. However, we focused on tf-idf primarily, as this yielded more interpretable results later in our analysis.

Lastly, we divided the merged data set into a train and test set using an 80/20 train-test split. We focused the bulk of our analysis on train data set and used the test data set later in the analysis for supervised learning and to examine the robustness of our results.

3.4 Latent Dirichlet Allocation

Latent Dirichlet allocation (Blei, Ng, Jordan, 2003) is a generative probabilistic model and a canonical method for dimension reduction in topic modeling [11]. It allows for unobserved topics to explain sets of observations, highlighting similarities and latent structures in data. More specifically, LDA is a mixed membership model, meaning that each document belongs to a distribution of topics as opposed to a single topic [11]. We implemented LDA using the SciKitLearn Python libraries [9].

In order to assess the robustness of our LDA models, we utilized log-likelihood and perplexity scores, from the requisite SciKitLearn methods, score and perplexity [12]. Larger values for the log-likelihood tend to indicate better success, and lower perplexity indicates better fit [12]. We normalize both scores using by the length of the data set.

We suspected that 2 topics would be appropriate for this analysis, as we were interested in seeing how well "rap" and "poetry" separated in topic space. However, we were cognizant of the fact that there may have been other interpretable latent topics in the dataset, so we applied LDA on the tf-idf representation of the data using 2, 3, and 4 for the number of topics and compared log-likelihood and perplexity scores in order to understand how best to proceed. As shown in Table 1, we found that 2 produced the highest log-likelihood score (-37.687) as well as the lowest perplexity score (0.0254). As a result, we use LDA with 2 topics for the remainder of the analysis; the composition of the latent topics are presented in the subsequent section. It should also be noted that we ran LDA with 2 topics for both the tf-idf representation and bag-of-words and found the normalized log-likelihood score for the former was -37.687 compared to -690.59, further reinforcing our choice to use the tf-idf representation for the bulk of the analysis.

Table 1: Log-Likelihood and Perplexity Scores

Number of Topics	Log-Likelihood	Perplexity
2	-37.687	0.0254
3	-38.359	0.0286
4	-38.542	0.0295

3.5 K-Means Clustering

K-Means clustering was performed on the topic proportions produced by LDA using SciKitLearn [9]. K-Means clustering is done by initializing k centroids, assigning each data point to the centroid closest to it, recalculating the centroid locations based on the mean of all data points corresponding to each centroid, and then repeating these steps until the difference in centroid locations between steps is less than a certain threshold [9]. While we assumed that the optimal number of clusters would correspond to the number of latent topics, we examined the results of the elbow method and silhouette method to validate our reasoning. Figure 11 (in Appendix) and Figures 5 through 8 show the results of the elbow and silhouette methods, respectively. In this rendering of the method, the elbow method calculates the sum of squared distances from each point, to its assigned center, also known as distortion, for each k [13]. As shown in Figure 11, the distortion value decreases significantly from k = 1 to k = 2 and does not increase after that.

With regards to the silhouette method, Figures 5 through 8 display the silhouette coefficient for each sample on a per-cluster basis as well as offers a visualization of which clusters are dense or not. The silhouette coefficient is calculated using the average intra-cluster distance and the average nearest-cluster distance, that is to say, the distance between a sample and the nearest cluster of which the sample is not a part [9]. The best value is 1, and the worst value is -1 [9]. In these plots, the x-axis shows silhouette coefficient values and and y-axis shows the cluster label. The dotted vertical line shows the average silhouette coefficient value. The silhouette coefficient is highest when k is 2, further reinforcing our selection of 2 for k in k-means. The silhouette scores are summarized in Table 2.

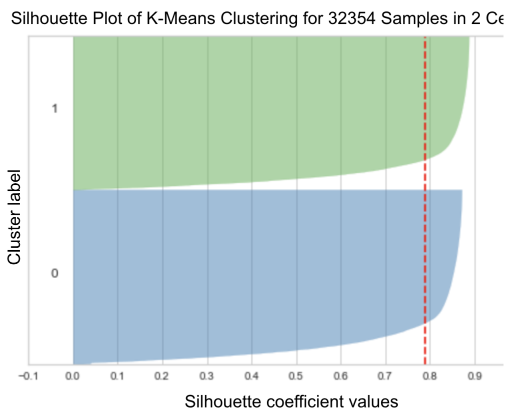

Figure 5: Silhouette Plot for 2 Clusters

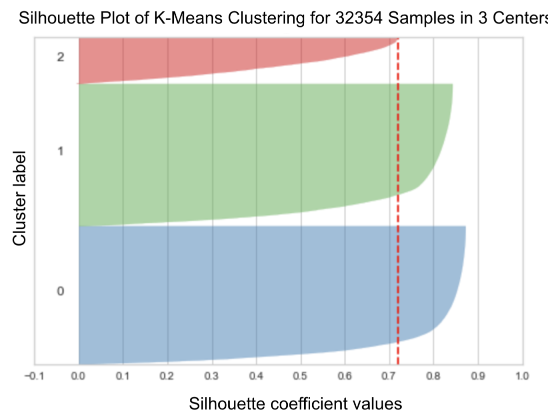

Figure 6: Silhouette Plot for 3 Clusters

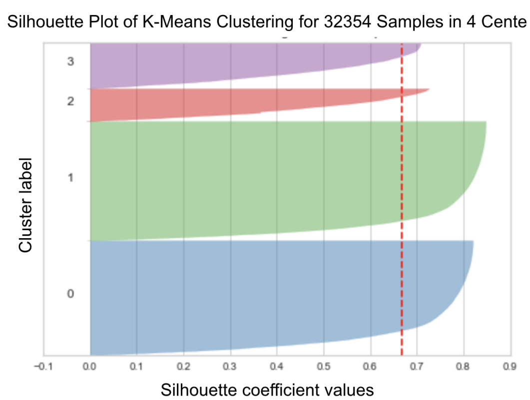

Figure 7: Silhouette Plot for 4 Clusters

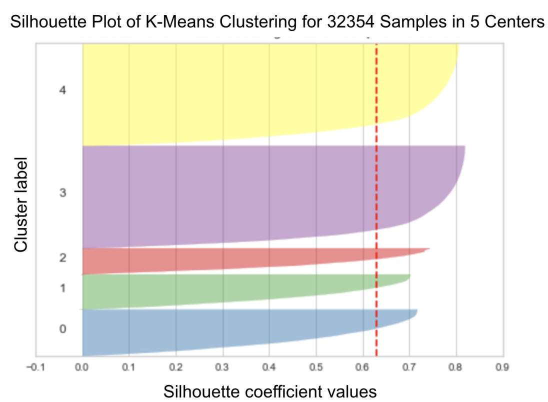

Figure 8: Silhouette Plot for 5 Clusters

Table 2: Silhouette Scores

k	**Silhouette Scores**
2	0.7892
3	0.7211
4	0.6691
5	0.6301

3.6 Classification

In part motivated by how well poetry and rap separated in topic space, we decided to perform binary classification and applied the LogisticRegression, SVC, and LinearSVC methods from the SciKitLearn library [9]. Each of these functions applies a supervised learning algorithm to predict an output value given the input [9]. These methods were each fit on the train data with the term frequency-inverse document frequency matrix as input, and assigned labels on the test data based on both the true labels of either poetry and rap (as designated in the data set) as well as the labels assigned by k-means clustering as output. After applying these classification methods to the entire tf-idf representation of the data, we applied the LogisticRegression method to an input of a 98% reduction in the feature space: the top 20 words found in LDA. This was performed in order to compare how well those specific words aid in classification – and thus how effectively split the LDA latent topic space was. Logistic regression was the chosen method for this extension due to it's relatively high precision-recall average score, and for ease of interpretability.

The purpose of running the initial classifiers was to further evaluate how well K-means clustering split the data between rap and poetry. Additionally, we hoped to compare classifiers and determine which one performed best on the data. All classification methods were evaluated using the score method in the SciKitLearn library [9]. This method returns the mean accuracy of the predicted values the model output. Further evaluation was done utilizing precision-recall curves, which plot proportions of true positive results over actual results versus true positive results over predicted results.

4 Results

4.1 Latent Dirichlet Allocation

In this analysis, we explored latent topics in the combined data set of contemporary poetry and rap by applying LDA in order to see how well the two classes separated in topic space. We ran LDA with two topics; this was motivated by the log-likelihood and perplexity results presented in the previous section. From the generated topics, we saw a divide, and so assigned each topic with a label we felt best described its content. We expected that the two classes would separate reasonably well but also wanted to see whether there would be other latent topics that highlighted any ambiguity between the two classes.

Table 3: LDA-Generated Latent Topics and Top 10 Words

Topic Number	Label	Tokens
1	rap	like, got, don, know, nigga, just, ain, cause, shit, niggas
2	poetry	like, love, light, night, day, eyes, world, life, away, long

After applying LDA to the tf-idf data, we analyzed the latent topics by looking at the top 20 words in each component, based on code adapted from Precept 9. Topic 1 was comprised of slang terms like "ain," "cause," "em," and "yo" as well as vulgar words. From this, we inferred that Topic 1 consisted primarily of words commonly used in rap. Topic 2 was comprised of words related to time, nature, and love such as "night," "time," "day", "world," "water," "sun," "love," and "god." Given that these words are commonly used in figurative language, we inferred that Topic 2 consisted of words commonly used in poetry. This reveals that the two classes – "rap" and "poetry" – separated well in the topic space. This was underscored by the fact that when we analyzed the top words for LDA run with three and four components, the resulting topics were not nearly as interpretable. Lastly, it was interesting that "like" was the top word in both topics, suggesting perhaps that both rap and poetry make heavy use of similes. This reinforces the idea mentioned previously that there exists structural overlap between the two classes. Table 3 presents the two topics with – for brevity's sake – the top 10 words.

4.2 K-Means Clustering

As mentioned above, we determined the optimal number of clusters using k-means on the topic proportions was 2. We found that centers for the two clusters were at (0.80860735, 0.19139265) and (0.2036359, 0.7963641), where the first component in the tuple is the proportion of Topic 2 (“poetry”) and the second Topic 1 (“rap”). These two centroids are designated by the white points in Figure 10. Using jitter, Figure 9 shows us the topic proportions for all of the poems and rap lyrics in our data set, with rap lyrics plotted in orange and poems in blue. Also using jitter, Figure 10 depicts the clustering of data points based on their topic proportions, which represents how we would have expected the data points to be labeled based on the results from LDA. In other words, documents with a higher topic 1 proportion were assigned to Cluster 1 and documents with a higher topic 2 proportion were assigned to Cluster 2. We found the two documents that were closest to the two centroids, as they serve as “archetypal” poems and raps according to their LDA topic proportions and K-Means. We include these examples in the appendix. We can clearly see that there are several rap data points in Figure 9 that were assigned to Cluster 2 according to their topic proportions, which indicates that these raps actually have a higher Topic 2 (“poetry”) proportion. We can also see that there were some (although considerably less) poems that were assigned to Cluster 1, indicating that these poems have a higher Topic 1 (“rap”) proportion. We examine some of these examples in more detail below.

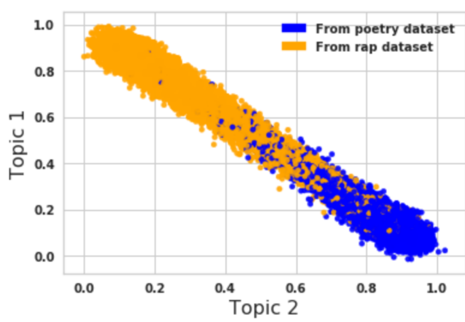

Figure 9: Topic Proportions and True Labels

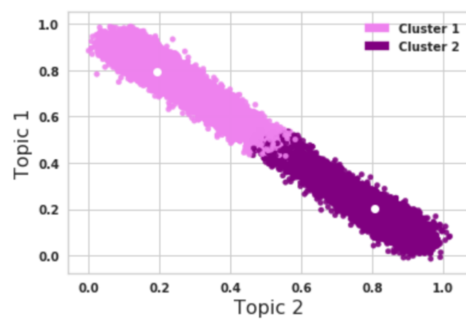

Figure 10: K-Means Clustering on Topic Proportions

4.2.1 Spotlights

There were 1,240 raps (3.65% of rap data set) that had a majority proportion of Topic 2, while there were only 13 poems (0.08% of poetry data set) that had a majority proportion of Topic 1. One example of a rap that had a majority proportion of Topic 2 is the song ”Where Were You” by Drake. Its Topic 1 proportion was 0.199, and its Topic 2 proportion was 0.801. A portion of the lyrics are as follows:

I can’t go to bed
I’m thinking about my goals
And how we used to say a goal is just a dream with a deadline
Where did the other half of my heart go?
Why am I in bed alone?
How come when I drive by, it looks like you are never home?
Do you even live there?
Or did you take a U-Haul?
Too far, move off and not even tell me that you’re gone?

Perhaps this rap was determined to have a majority Topic 2 proportion since it doesn’t contain slang or profanity as was demonstrated in the Topic 1’s top words. In addition, this rap song is more “lyrical” in nature and makes use of imagery and figurative language characteristic of poetry.

An example of a poem that had a majority proportion of Topic 1 is ”Resume” by Dorothy Parker. Its Topic 1 proportion was 0.773 and its Topic 2 Proportion was 0.227. The poem is as follows:

Razors pain you;
Rivers are damp;
Acids stain you;
And drugs cause cramp.
Guns arent lawful;
Nooses give;
Gas smells awful;
You might as well live.

This poems contains words such as “guns” and “drugs,” which relates to themes commonly found in rap, explain why its topic distribution misaligned with its actual class. We include excerpts of other examples in our appendix and, broadly speaking, find that poems with end rhyme and repetition were, in some cases, clustered with mostly raps and raps with more figurative language and imagery and fewer slang words were often assigned to the cluster containing mostly poems. These results could potentially back-up a claim that rap can be considered as a subset of poetry, but that not all poetry can be considered as rap.

4.3 Classification

After applying multiple classification methods, we found a notable decrease in accuracy scores when using cluster labels, compared to the actual labels, as expected. These results were utilized to evaluate how well the unsupervised learning methods applied split the data into the two classes. The highest score was achieved with logistic regression on the minimized feature space from the latent topics in LDA, which was done to examine the effectiveness of classifying the data with a small number of key features, while the highest average precision-recall score was in logistic regression on the entire tf-idf feature space. The scores and average precision-recall scores (PR) are shown below in Table 4, where LR stands for logistic regression, SVC for support vector classifier, LSVC for linear support vector classifier, and LR(on LDA) for logistic regression utilizing only the features from the key words from the latent topics found in LDA. In the table, actual refers to classifiers fit on the actual labels for rap and poetry, and cluster indicates classifiers fit on the labels created by k-means clustering. The plotted precision-recall curves provide a useful visualization of these results, and is visible in Figure 12, located in the appendix. The results indicate future applications of logistic regression could offer interesting insights, and that the LDA topic words are an effective feature space, despite the small size, which could be applied to other data for classification purposes.

Table 4: Accuracy and Precision-Recall scores for Classifiers

	LR		**SVC**		**LSVC**		**LR (on LDA)**
	actual	cluster	actual	cluster	actual	cluster	actual
Score	.571	.524	.619	.560	.604	.537	.645
PR	.638	.551	.607	.545	.618	.565	.607

5 Discussion and Conclusion

In this project, we sought to understand better the distinction between rap and modern poetry via unsupervised and supervised learning. We used Latent Dirichlet Allocation analysis with two topics to gauge whether there would be a clean split in topics between the two categories. While we predicted that there would be a somewhat clear split, we were surprised by the extent to which the topics were distinguishable. This is important as LDA shows us that the topics in the two categories are separable to a significant extent. We also extracted the topic proportions for each sample from the LDA analysis to perform K-Means Clustering, with k=2. The clusters then represented what each sample would have been labeled as according to its topic proportions. By comparing the cluster assignments of each sample to its true labels, we were able to determine that there were many more raps that had a high poetry topic proportion than the other way around. This is especially significant as it indicates that rap can have a certain ambiguity in terms of identity that was not captured explicitly in the LDA analysis. Moreover, it reinforces the idea presented previously that there exist cases in which the distinction between the two art forms is not so clear and highlights why this might be the case. Lastly, we were interested in performing supervised learning to see what features were most predictive of the category. We found that using logistic regression solely on the LDA topic words produced relatively strong results, indicating that those words are the most predictive of category, as we expected. There are several ways that we could extend and improve this project. Since there are several types of rap within the broader category, we could delve deeper into which types of rap are most and least distinct from modern poetry using potentially more topics. In addition, it would be interesting to track key artists and understand how their work evolves with respect to how poetic their lyrics are, as well as how their rap differs from those of their peers. More sophisticated unsupervised learning methods, such as Dirichlet process mixture models or dynamic topic models, could also provide an interesting angle for comparison and analysis.

6 Appendix

6.1 Additional Plots

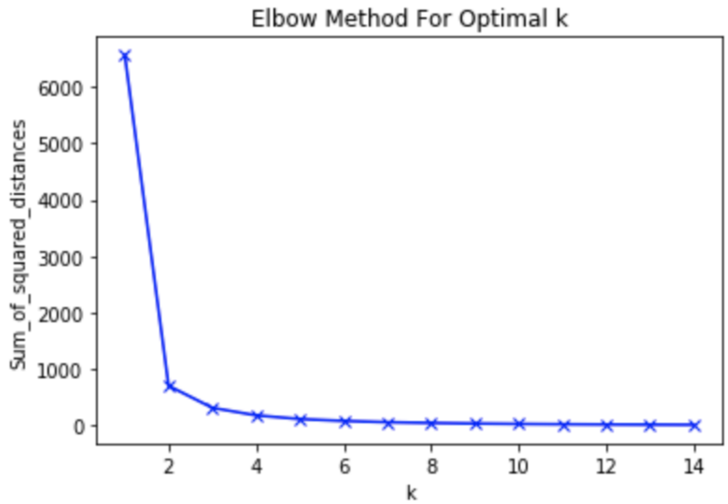

Figure 11: Elbow Plot Showing the Optimal K

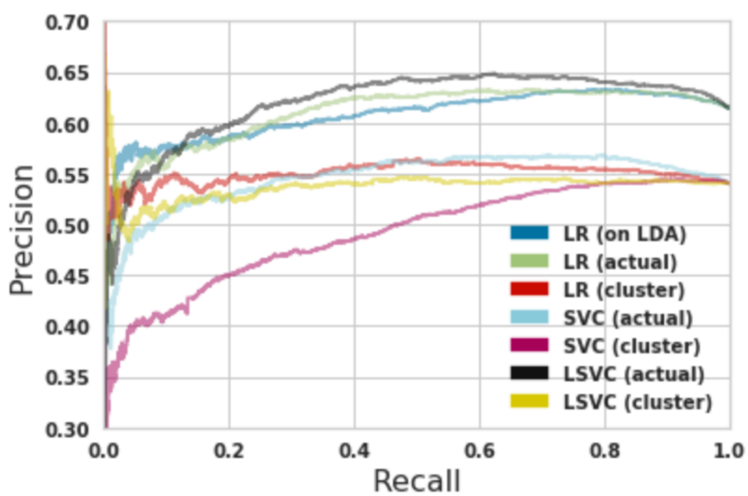

Figure 12: Precision-Recall Curves for Classification Methods

6.2 Additional Rap and Poetry Spotlights

“Cliff Hanger” by Blackalicious
Class: Rap
Topic Distribution – Topic 1: 0.155 and Topic 2: 0.845

It was a Thursday I stepped to the club, there she was
Draped in a burgundy dress, silhouetted superbly
Curved indeed, she started staring in my direction
Luring and urging seductively, body language instructing me

I offered her a drink instead she took a cup of tea
And then I lied about my luxury lifestyle
I said "So who you be? You looking kind of tight, gal
I would love to see you in a nightgown"

"Drag On" by Mutant Class – Rap
Topic Distribution – Topic 1: 0.104 and Topic 2: 0.896

How quickly seconds fly
Pendulum in my hands
A mutant
In hopeless desperation
I know that's my day
The day when my world will be gone

"Mix a Pancake" by Christina Rossetti
Class – Poem
Topic Distribution – Topic 1: 0.802 and Topic 2: 0.197

Mix a pancake,
Stir a pancake,
Pop it in the pan;
Fry the pancake,
Toss the pancake
Catch it if you can.

References

[1] Mattix, M. Is Rap Poetry?. *The American Conservative*. https://www.theamericanconservative.com/prufrock/is-rap-poetry/.

[2] rocawear. JAY-Z "Rap is Poetry". *YouTube*. https://www.youtube.com/watch?v=HXR-ohNo3Ao.

[3] Coscarelli, J. Kendrick Lamar Wins Pulitzer in Big Moment for Hip-Hop. *The New York Times*. 2018. https://www.nytimes.com/2018/04/16/arts/music/kendrick-lamar-pulitzer-prize-damn.html.

[4] Canini K, Shi L, Griffiths T. Online Inference of Topics with Latent Dirichlet Allocation. In: Artificial Intelligence and Statistics; 2009. p. 65-72.

[5] Rhody, L. Topic Modeling and Figurative Language. In: Journal of Digital Humanities; 2012. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

[6] Guan P, Wang Y, Chen B, Fu Z. K-Means Document Clustering Based on Latent Dirichlet Allocation. In: Western Decision Sciences Institute 2016 Proceedings; 2016. http://wdsinet.org/Annual_Meetings/2016_Proceedings/papers/Paper45.pdf.

[7] Mishra, G. 380,000 Lyrics from MetroLyrics. Kaggle; 2017. https://www.kaggle.com/ultrajack/modern-renaissance-poetry.

[8] ultra-jack. Poems from PoetryFoundation.org: Modern and Renaissance Poetry for Classification Exercises. Kaggle; 2017. https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics.

[9] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learnning in Python. Journal of Machine Learning Research. 2011; 12; p. 2825–2830. https://scikit-learn.org/stable/.

[10] Calderon P. Bag of Words and Tf-idf Explained. Data Meets Media; 2017. http://datameetsmedia.com/bag-of-words-tf-idf-explained/.

[11] Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. In: Journal of Machine Learning Research, 3; 2003. p. 993-1022.

[12] LDA in Python, How to Grid Search the Best Topic Models.https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/

[13] Bengfort B, Danielsen N, Bilbro R, Gray L, McIntyre K, Richardson G, Miller T, Mayfield G, Schafer P, Keung J. Yellowbrick: Machine Learning Visualization. https://www.scikit-yb.org/en/latest/index.html