A Lyrical Analysis of the Bee Gees' Discography

Maika Kinoshita Nebgen

Data and Social Media Analysis
kinoshitamaika@fuji.waseda.jp

30 January, 2025

Abstract

Known for their multi-genre reputation, an attempt was made to discover to what extent the Bee Gees' lyricism affects their lengthy success in the music industry. By examining the lyrics they use in their songs, a hypothesis was made that the repetition of their song lyrics would be a great indicator to their success, but they would have a unique approach to what lyrics would be used in comparison to the top songs in each decade. This paper discusses the lyrical analysis of the Bee Gees' discography and compares it to other top songs from the Billboard's Top 100 Hits from the 1960s to the 2000s. The resulting findings failed to depict a resourceful reliance on Latent Dirichlet Allocation (LDA) and BERTopic modeling used for determining semantic patterns in their discography; however, the utilization of Term Frequency-Inverse Document Frequency (TF-IDF) resulted in some similar term frequencies between the Bee Gees and other top hits, showing their correlation to trending lyrical writing for each genre involved, and solidified the repetitiveness that their songs obtain.

1 Introduction

The Bee Gees, a widely popular and successful band of brothers, have retained their worldwide renounce since the 1960s. The Bee Gees started their career as children, first performing in local theaters in the Isle of Man (Eder.) Later having moved to Australia, they were first known as the Brothers Gibb, until they gained popularity through television and changed their name to the Bee Gees. However, the emergence into stardom didn't occur until manager Robert Stigwood—notorious for working with Eric Clapton and the correlating band Cream—discovered them and signed them onto his record label RSO Records (Ruhlmann.) The Bee Gees maintained a relatively steady career until they disbanded temporarily due to complications over their album Odessa. Afterwards, with the release of Cucumber Castle by only two of the three brothers—Barry and Maurice Gibb—the brothers reunited with their third, Robin, and continued to build their reputation. It wasn't until the disco era that their album Spirits Having Flown took off their career and gained them national recognition (Eder.) Afterwards, the Bee gees continued to make history, with their emergence in falsetto singing in Main Course and the first successful creation of a drum loop in music production in the soundtrack for the film "Saturday Night Fever" (Marshall.) The Bee Gees continued with their incredible reputation to release a multitude of albums and write songs for other elite artists, including Otis Redding and Celine Dion (Marshall.)

To further investigate why their success has been maintained throughout decades of differing music genres and generations, a lyrical analysis of their most prominent albums has been conducted. Moreover, to maintain a threshold of what success entails, the Bee Gees' lyrics have been compared to the Billboard's Top 100 Hits for the corresponding decades.

This lyrical analysis was conducted to see if there was any correlation between the Bee Gees' lyricism and their lifelong success.

The original idea was to discover related topics within the lyrics of each of these albums to identify relative similarities to trends of each correlating album's genre at the time. Through this, the possibility of a lyrical structure or algorithm could be identified to understand the maintenance of success that the Bee Gees had for each generation of music released. Whether the Bee Gees succumbed to generational trends or popular themes in each genre to maintain their success via lyrical expression was attempted. Contrastingly, if their lyrics were uniquely outstanding from popular hits or one-hit wonders of each aspiring genre, was it this integrity to maintain their individuality amongst a booming genre that allowed them their worldwide status? To achieve this goal, BERTopic modeling was attempted, both with and without preprocessing steps to ensure peak performance. Unfortunately, the results deemed insufficient for a reliable analysis of correlating the Bee Gees' lyricism to the direction of their career. Upon further examination, another method was constructed to compare lyrics more understandably and with less intervals for misinterpretation. By utilizing TF-IDF and term frequency calculations, a comparison was made between the top words used in the Bee Gees' songs with the Billboard Top 100 Hits from the 1960s to the early 2000s (Hsu and Xu.) By looking at the most words used per album and comparing them to the top hits, one could infer the similarities that influenced the direction of the Bee Gees within each genre and consider their reputation of the corresponding time to formulate an in-depth analysis of not only why they have remained successful over the years, but when exactly did they seem to gain the most popularity or admiration.

By considering the musical career path, genres, top hits, and lyrical choices of the Bee Gees, we can gain a more concise understanding of why they have retained their success for such a notoriously long era. Furthermore, we can distinguish to what extent of their musical reputation was reliant on the lyrics of their songs and why they caught the attention of a global audience.

2 Background

Multiple projects consisting of lyrical analyses and genre comparisons have been conducted. The most prominent work that will be consistently referred to in this project is the work of two students at Brown University: Hsu and Xu. By extracting data from the Billboard Year-End Top 100 for the years from 1964 to 2015, they were able to come to conclusions about patterns in the lyrics of songs of differing decades, categorizing words by uniqueness, bigrams, trigrams, top 15 words used per song, and sentimental words per decade. Their results led to the conclusion that each decade showed a consistent use of the words "love," "girl," and "time." Additionally, there was a gradual increase in profanity and simplicity in lyrics as the uniqueness went up until the 2010s, where it dropped significantly. The simplicity of the lyrics may be due to the repetitiveness for each song, showing a pattern in top 100 songs that may be associated to the results of the Bee Gees' discography as well.

Another study was done to further examine the growing simplicity in songs, and why they were gaining success in the music industry (Varnum et al. (2021).) The largest and most prominent factor found was the number of songs released each year (Rigg (2021).) Due to the influx of song releases, audiences may find it more difficult to consume each song as the number of available songs increases. Therefore, decreasing the complexity of each song showed more successful results for the artist in competition with a multitude of other releasing songs, especially when they obtain repetitive motifs or lyrics that allow for easier memorization of the listener (Varnum et al. (2021).) The level of simplicity for each song was calculated by measuring the compressibility for each song (Varnum et al. (2021).)

3 Data

Twenty-one of the Bee Gees' albums have been chosen for this lyrical analysis for two reasons: the first being that these are the most notable and popular albums from the Bee Gees' discography and of their musical career, and the second reason being that some albums that were also deemed significant to their career were unable to correspond to the data requirements during the data collection process due to technical issues. Albums that are considered anthologies and "Best Hits" albums were also not considered to reduce the redundancy in data.

The albums chosen are as follows in chronological order of release date. Songs that don't include lyrics, such as instrumental songs, and collaborative albums of songs by other artists have been excluded from the data below and are marked with an asterisk. The file name used in the data are also shown:

Album Title	Release Year	.txt File Name
"The Bee Gees Sing and Play 14 Barry Gibb Songs"	1965	barry_gibb_songs
"Spicks and Specks"	1966	spicks_and_specks
"Turn Around, Look at Us"	1967	turn_around_look_at_us
"Bee Gees' 1st"	1967	bee_gees_1st
"Horizontal"	1968	horizontal
"Idea"	1968	idea
"Odessa"	1969	odessa
"Cucumber Castle"	1970	cucumber_castle
"2 Years On"	1970	two_years_on
"Trafalgar"	1971	trafalgar
"To Whom It May Concern"	1972	to_whom_it_may_concern
"Mr. Natural"	1974	mr_natural
"Main Course"	1975	main_course
"Children of the World"	1976	children_of_the_world
"Saturday Night Fever"	1977	saturday_night_fever
"Spirits Having Flown"	1979	spirits_having_flown
"E.S.P"	1987	esp
"One"	1989	one
"High Civilization"	1991	high_civilization
"Size Isn't Everything"	1993	size_isnt_everything
"This is Where I Came In"	2001	this_is_where_i_came_in

The lyrics from each album were extracted using a Genius API (Genius) and relocated into .txt files for easy manipulation. Each file represents one album, and each line within the files represent one song from the corresponding album. A total of 249 songs were extracted from the Genius website to analyze lyric topics and word frequencies using BERTopic modeling, LDA topic modeling, and TF-IDF detection. Data on the Billboard Top 100 Hits of the Hsu and Xu project will be used for term frequency comparison of each corresponding genre for the Bee Gees' discography (Hsu and Xu.)

4 Methods

4.1 Data Collection

To collect the lyrics for each song, a loop was created using the API token to search through each URL for each song of each album. Assistance in curating the code was derived from: Nick Pai, Juico Bowley, Rithesh Sreenivasan, MachineLearningPlus,

Koray Tugberk Gübür, GeeksforGeeks, and Heinz-Alexander Fuetterer.

4.2 LDA Topic Modeling

The first attempt to depict patterns in topics of each album was used with LDA topic modeling. In order to conduct LDA topic modeling, preprocessing was conducted by removing stopwords, removing any unnecessary characters and punctuation, tokenizing, and lemmatizing each album file. Then, each file was categorized into five topics each to determine any relating patterns in their lyrics.

4.3 BERTopic Modeling

The next attempt to distinguish topics within each album was done using BERTopic modeling, due to its exclusion of the preprocessing steps. This showed potential because the lyrics didn't have to be cleaned, therefore maintaining the original quality of the lyrics and possibly revealing more information than the attempt with LDA Topic Modeling. Using BERTopic also accustomed to the data rather than having to manually adjust the data with LDA; the topics were chosen by BERTopic, whereas when using LDA, the number of topics had to be chosen before results could appear.

4.4 TF-IDF Acquisition

The third and most successful attempt was done with TF-IDF acquisition. Especially due to the unsubstantial data derived from the previous two attempts, a change in direction had to be made from discovering the semantics of each album to discovering the relatability of each album to its corresponding popular hits of that decade. This was done with different attempts including and excluding stopwords and customizing the search for term frequency for each album file.

5 Results

5.1 LDA Topic Modeling

This attempt proved inconclusive to depict any substantial data on topics due to the removal of most of the lyrics during the preprocessing steps: due to the simplicity of the lyrics, many of the words may have been considered as stopwords or were cleaned inaccurately. When checking the data after preprocessing, more than half of the data had been excluded, so the visualizations of each topic were not reflective of each album.

5.2 BERTopic Modeling

Four separate attempts were made using BERTopic modeling:

5.2.1 No preprocessing for all album files

No preprocessing or customization led to a result of the -1 topic, or unique topic, being assigned to all albums and their data. This led to inconclusive and unusable data.

5.2.2 Assigning outliers to attempt collection of more than 1 topic per album

By assigning outliers to the data, an attempt to get more than one topic per album was made. However, this led to an error resulting in insufficient data for each outlier. In other words, there was not enough data to create different outliers, and more than one topic could not be distinguished.

5.2.3 Using the clustering algorithm kmeans to prevent outliers and enforce topic creation

The KMeans clustering algorithm from SciKit Learn was utilized in an attempt to enforce a more concise topic categorization. While a low number of topics were distinguished, they were still insufficient for any data analysis or comparison to each album and each hit of the decades. Each topic's top five categories were all commonly used words, such as "you," "and," "the," etc. This has shown the most potential, however, so a final attempt with the exclusion of stopwords was done to see if the topics would change with more unique data.

5.2.4 Continued using kmeans but with stopword exclusion to obtain a more substantial data output

Unfortunately, the exclusion of stopwords didn't differentiate from the third attempt using BERTopic in 5.2.3. The top topics still showed the same commonly used words, so an overall different approach was used using TF-IDF.

5.3 TF-IDF Acquisition

By using TF-IDF, the most conclusive and understandable data was shown and comparable to other Billboard Top Hits. Four attempts were also made here to improve and customize the search for data:

5.3.1 Included stopwords

When including stopwords in the visualization graphic, the amount of words shown was much larger than the scale of the graph and comparison for each document's TF-IDF score, so the next step would be to exclude stopwords to lessen the term frequencies and therefore make the visualization legible and comprehensible.

5.3.2 Excluded stopwords

Excluding stopwords and limiting the visualization to the top ten words for each document (album) showed much more understandable data; however, the data still had terms like "ve," "ll," and "ain," as some of the top ten words. This was most likely from contraction words, such as "I've,", "I'll," or "ain't."

5.3.3 Excluded stopwords and words less than 3 letters in length

The next step was to adjust and customize the search of term frequencies. Seeing as majority of the top ten words were words longer than 3 letters and most of the ineligible words were less than 3 letters long, this step seemed the most reasonable to take. The data was finally usable for comparing to other top songs of the decades and some analysis was plausible for each Bee Gees album.

5.3.4 Improvements to visualization tactics

Rather than showing the TF-IDF scores of each word, the term frequency was calculated instead to be able to compare data with the Billboard Top 100 Hits collected by Hsu and Xu. Additionally, each word in the bar charts were associated with one color throughout each album, but some colors are still overlapping due to the limitations in color availability.

6 Analysis and Discussion

By looking at attempts 5.3.3 and 5.3.4 of the TF-IDF method, the data seems similar to Hsu and Xu's findings of the top terms used per decade: majority of the Bee Gees' albums—with exceptions to Two Years On—showed that the word "love" was of the top ten most frequent words used in their lyrics. Moreover, "love" was the most frequently used word for half of the albums. It is important to note, also, that "don" was a frequent term in some albums, and it is estimated that this is a fragment of the word "don't" that was accidentally cut off.

Another observation is that majority of the most frequent words are the same as the song titles for each album. For example, in the album "Children of the World," the song "Lovers," has its name as the second-most frequent word throughout the entire album. This is most likely due to the repetitive enunciation of the word throughout the song, rather than its mentioning spread out within the entire album. This repetition of the song's title is also noticeable in the albums "Odessa," with the song "Whisper Whisper," (Figure 7) the album "Spirits Having Flown," with the song "Living Together,"—although both "livin" and "living" appear in the top ten terms for this model (Figure 16)—and the album "Horizontal," with the song "Harry Braff" (Figure 5.)

7 Limitations

When measuring the musical success of an artist, not only the lyrics should be taken into consideration. Their fan base and public representation via interviews, talk shows, and other methods of publicity should also be considered as to how their success fluctuates. In addition to the lyrics of their songs, their instrumentation—especially when making comparisons by genre or trending decade—should be thoroughly examined. Genres are primarily an indicator of a shift in social trends, but more so in instrumentation and timbre of music. Therefore, background vocals, music notation and music theory, and what instruments are used at the time need to be more strictly associated to musical success than any other aspect of an artists' representation.

In terms of the data, attempting to use topic modeling and semantic categorization of lyrics will often result to incomprehensible data. This is because methods such as topic modeling or sentiment analysis find it difficult to take anything other than the literal interpretations of text. Poetic or artistic writing such as lyrics pose meanings that vary between each listener, making it a difficult resource for a computer to analyze in a manner that humans would agree with.

8 Conclusion and Future Work

While the data for the Bee Gees' lyrics weren't directly reflective of the Billboard Top Hits, there was a major similarity between the most commonly used word "love" throughout almost all of their albums. Another significant point was the repetitiveness of song titles used as a frequent term due to the notorious repetition that the Bee Gees use in their song lyrics.

In order to further analyze the discography of the Bee Gees and find more concise data, using machine learning algorithms to visualize music notation and compare their songs with other genres would reveal more conclusive information. Some works on visualizing music notation using topic modeling and machine learning have already been done by Spiliopoulou and Storkey, Spiliopoulou and Storkey (2011), Panda et al. (2021), and Shalit et al. (2013). Perhaps the combination of musical and lyrical notation may provide a more in-depth analysis of the Bee Gees discography and widen the means of research to be done towards this band of brothers.

9 Code

My code can be looked at here: GitHub Repository: DSMA-Final-Project. "genius" refers to **Section 3** on data collection as well as the process for LDA Topic Modeling in **Section 4.2**. "genius2" refers to the process for BERTopic Modeling in **Section 4.3**. "genius3" is for TF-IDF analysis in **Section 4.4**.

References

B. Eder. Bee gees biography. *AllMusic*. URL `https://www.allmusic.com/artist/bee-gees-mn0000043714#biography`.

GeeksforGeeks. Understanding tf-idf (term frequency-inverse document frequency). URL `https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/`.

Genius. URL `https://docs.genius.com/#/getting-started-h1`. Genius API acquisition.

Heinz-Alexander Fuetterer. URL `https://github.com/MaartenGr/BERTopic.git`.

E. Hsu and H. Xu. Analysis of billboard's top 100 songs and lyrics (1964-2015). Brown University. URL `https://cs.brown.edu/courses/cs100/students/project11/`.

Juico Bowley. URL `https://github.com/Juicob/selenium-genius-lyrics-scraper.git`.

Koray Tugberk Gübür. How to use bertopic for topic modeling and content analysis? URL `https://www.holisticseo.digital/python-seo/topic-modeling/`.

MachineLearningPlus. Topic modeling with gensim (python). URL `https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#2prerequisitesdownloadnltkstopwordsandspacymodelforlemmatization`.

Marshall. The bee gees: How can you mend a broken heart. URL `https://www.imdb.com/title/tt9850386/`. Bee Gees documentary, HBO Max.

Nick Pai. How to scrape song lyrics: A gentle tutorial. URL `https://medium.com/analytics-vidhya/how-to-scrape-song-lyrics-a-gentle-python-tutorial-5b1d4ab351d2`.

S. Panda, V. Namboodiri, and S. T. Roy. Visualizing music genres using a topic model. *ResearchGate*, 2021. URL https://www.researchgate.net/publication/349703907_Visualizing_Music_Genres_using_a_Topic_Model.

C. Rigg. Newer generations prefer simpler song lyrics. *Psypost*, 2021. URL https://www.psypost.org/newer-generations-prefer-simpler-song-lyrics/.

Rithesh Sreenivasan. URL https://github.com/rsreetech/LDATopicModelling.git.

W. Ruhlmann. Robert stigwood biography. *AllMusic*. URL https://www.allmusic.com/artist/robert-stigwood-mn0000288781.

U. Shalit, D. Weinshall, and G. Chechik. Modeling musical influence with topic models. *ResearchGate*, 2013. URL https://www.researchgate.net/publication/290080267_Modeling_musical_influence_with_topic_models.

A. Spiliopoulou and A. Storkey. A topic model for melodic sequences. URL https://icml.cc/2012/papers/585.pdf.

A. Spiliopoulou and A. Storkey. Comparing probabilistic models for melodic sequences. *School of Informatics, University of Edinburgh, United Kingdom*, 2011. URL https://homepages.inf.ed.ac.uk/amos/publications/SpiliopoulouStorkey2011MelodicSequences.pdf.

M. E. W. Varnum, J. A. Krems, C. Morris, A. Wormley, and I. Grossmann. Why are song lyrics becoming simpler? a time series analysis of lyrical complexity in six decades of american popular music. *PLOS One*, 2021. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0244576.

10 Appendix

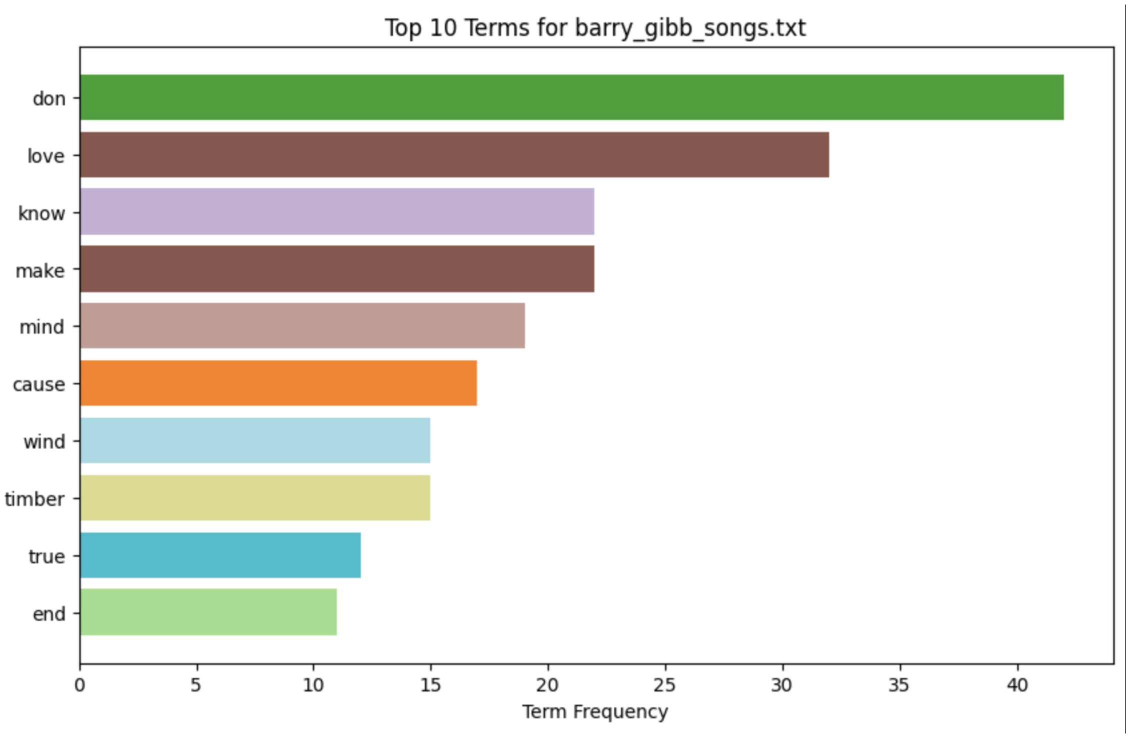

Figure 1: Top 10 Words in "The Bee Gees Sing and Play 14 Barry Gibb Songs"

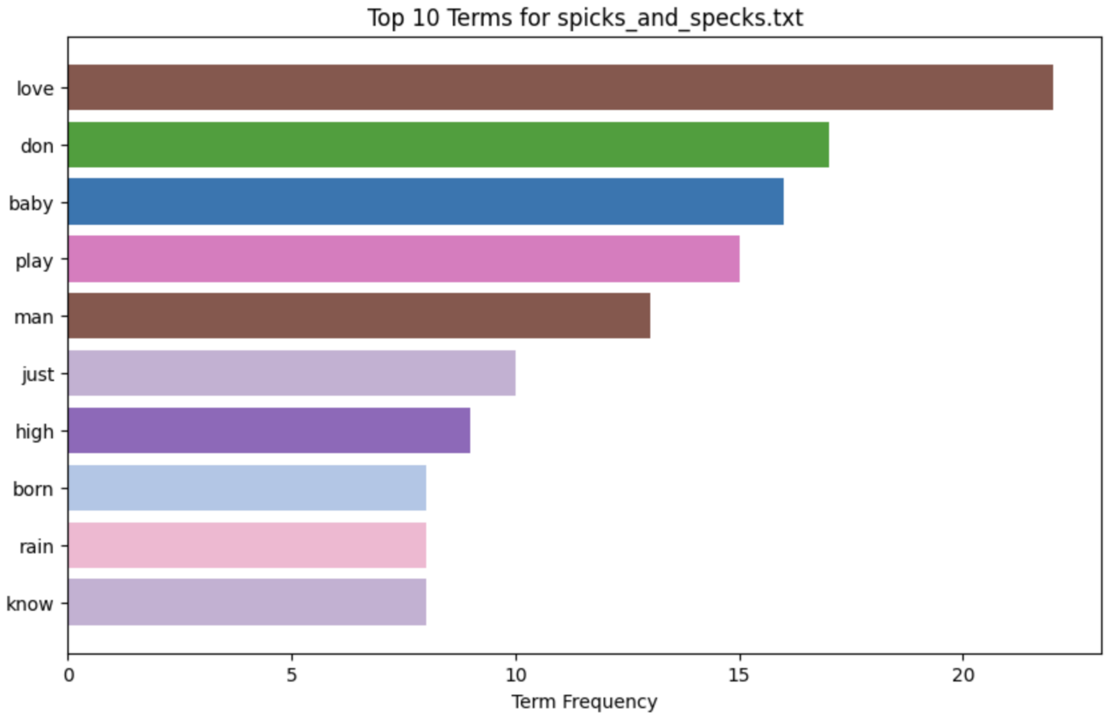

Figure 2: Top 10 Words in "Spicks and Specks"

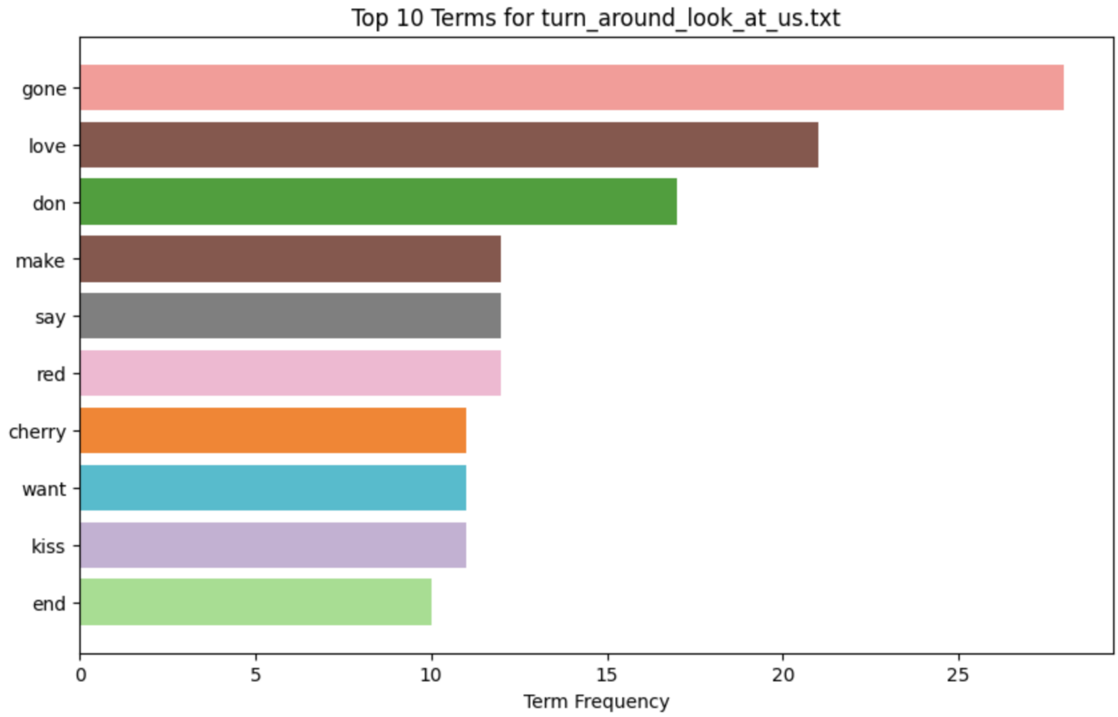

Figure 3: Top 10 Words in "Turn Around, Look at Us"

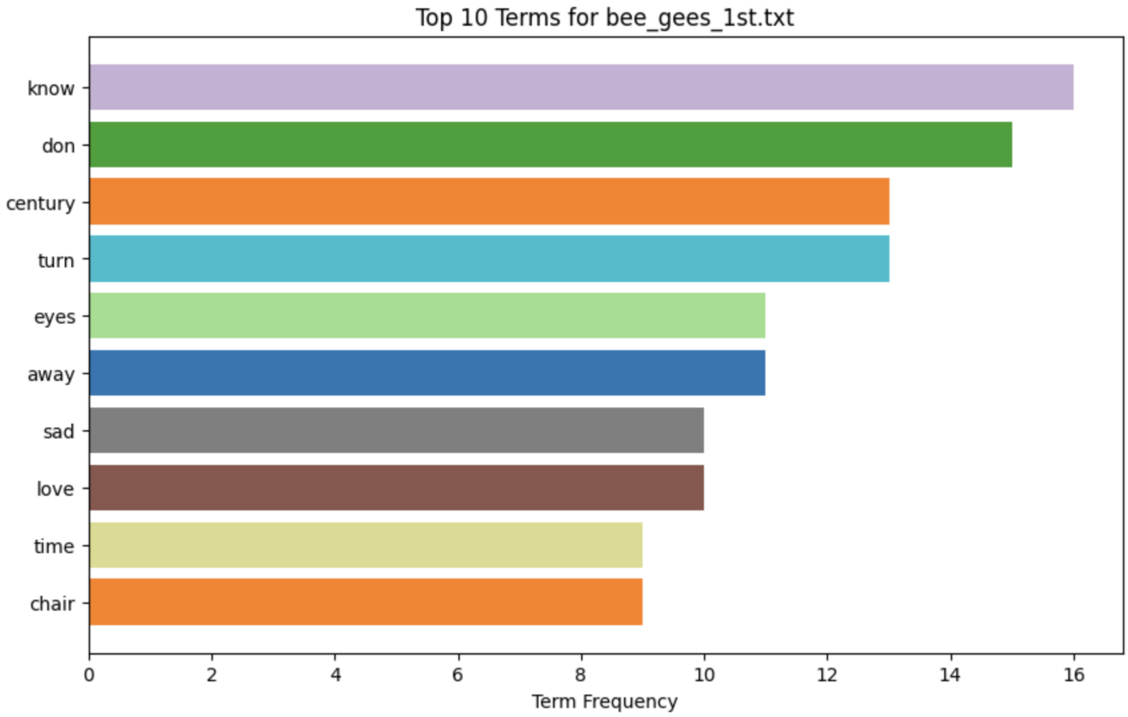

Figure 4: Top 10 Words in "Bee Gees' 1st"

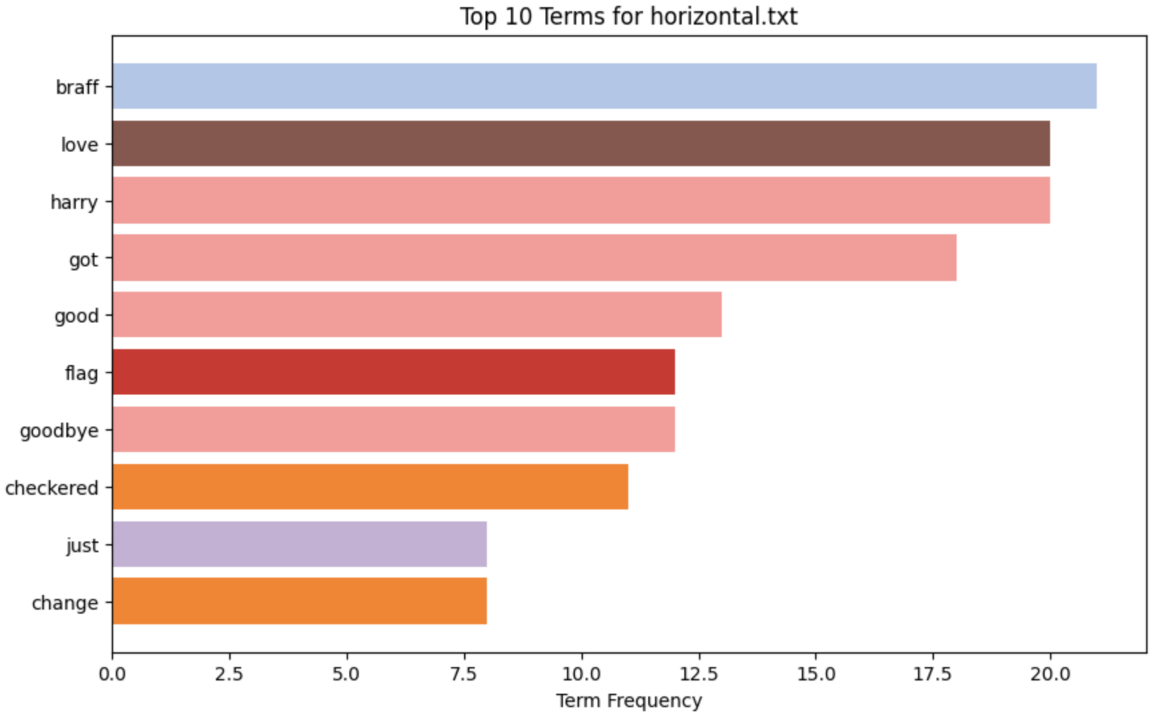

Figure 5: Top 10 Words in "Horizontal"

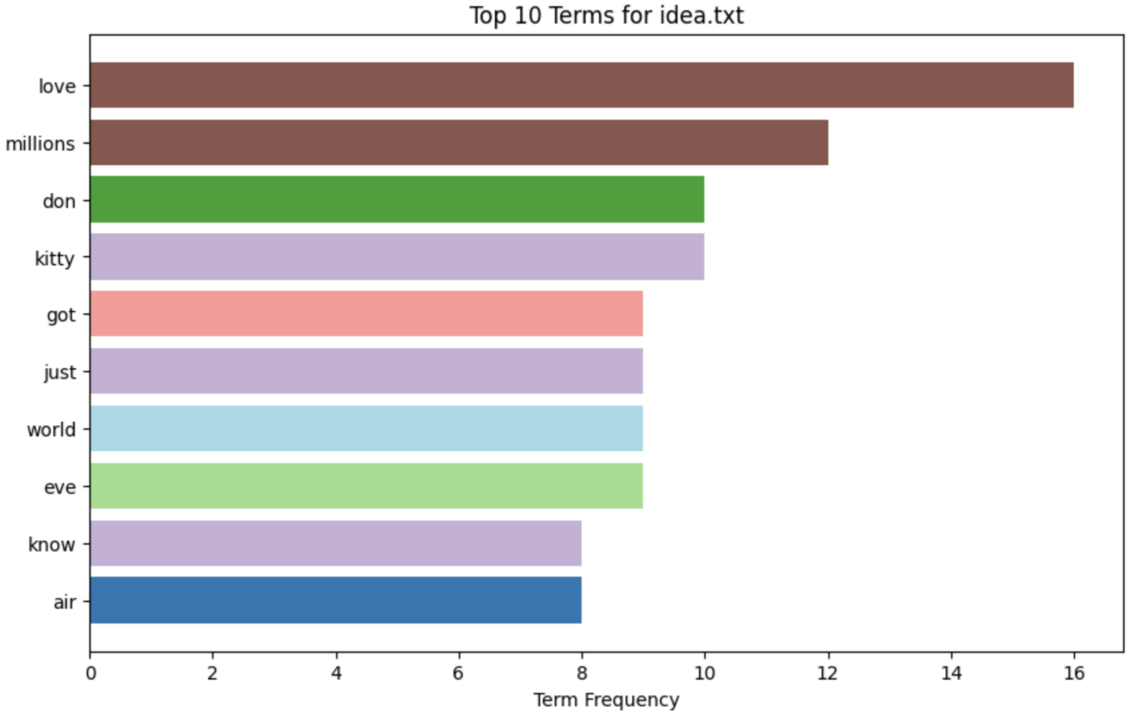

Figure 6: Top 10 Words in "Idea"

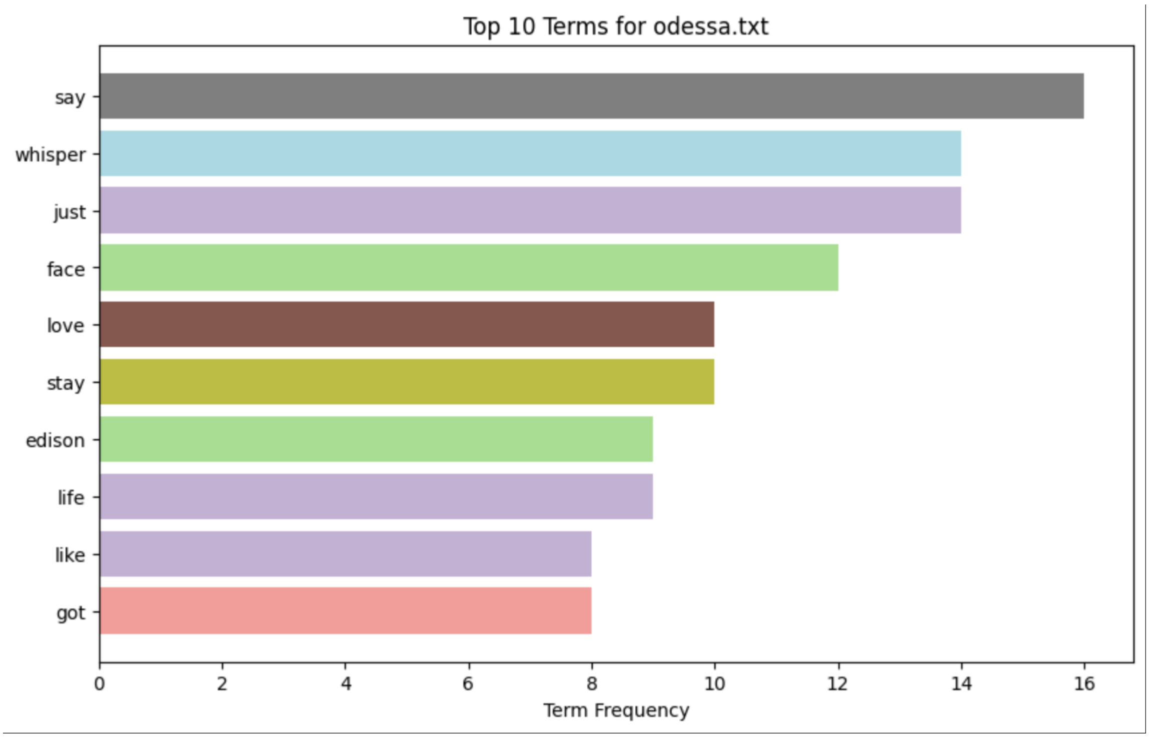

Figure 7: Top 10 Words in "Odessa"

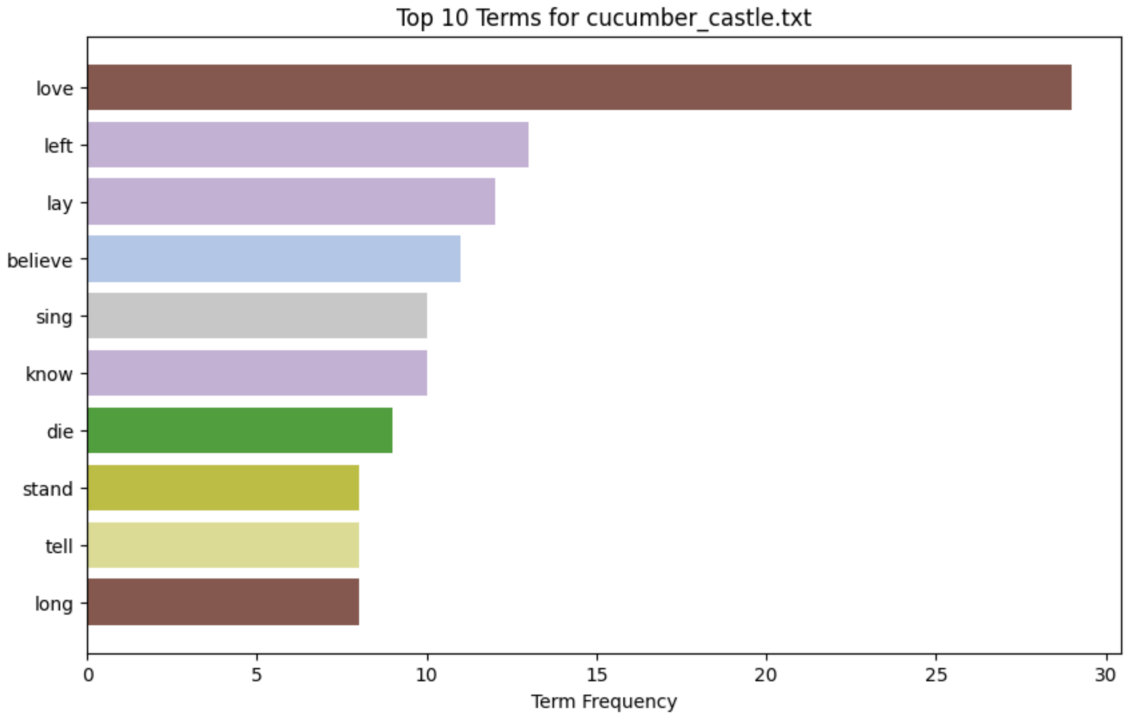

Figure 8: Top 10 Words in "Cucumber Castle"

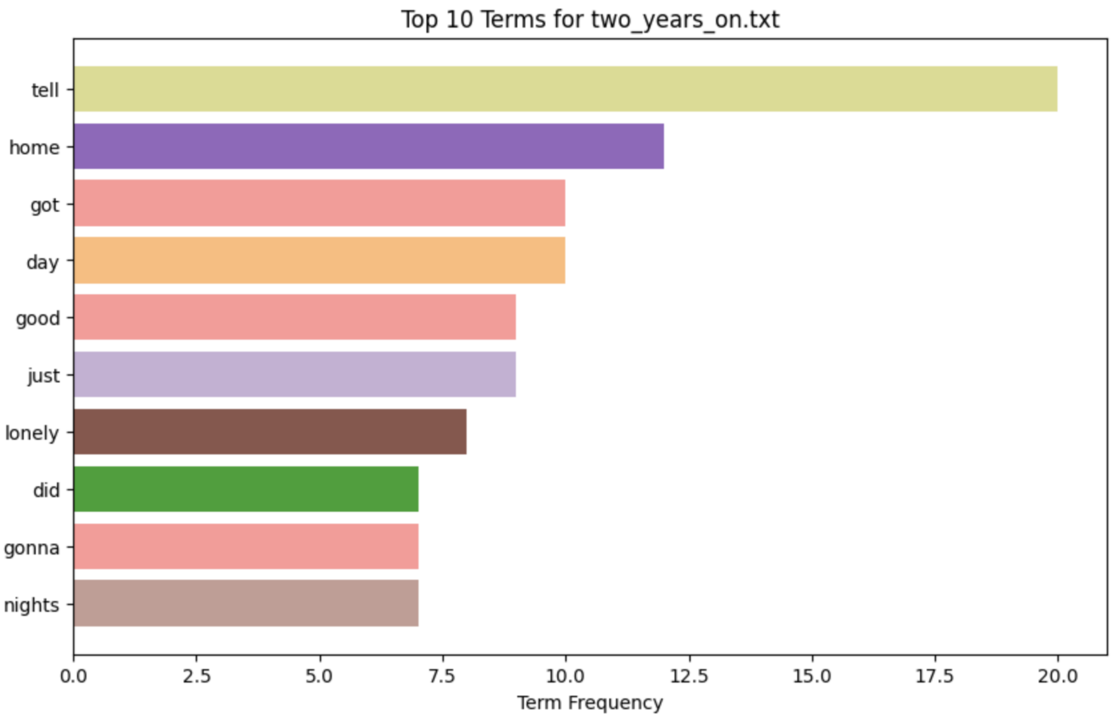

Figure 9: Top 10 Words in "2 Years On"

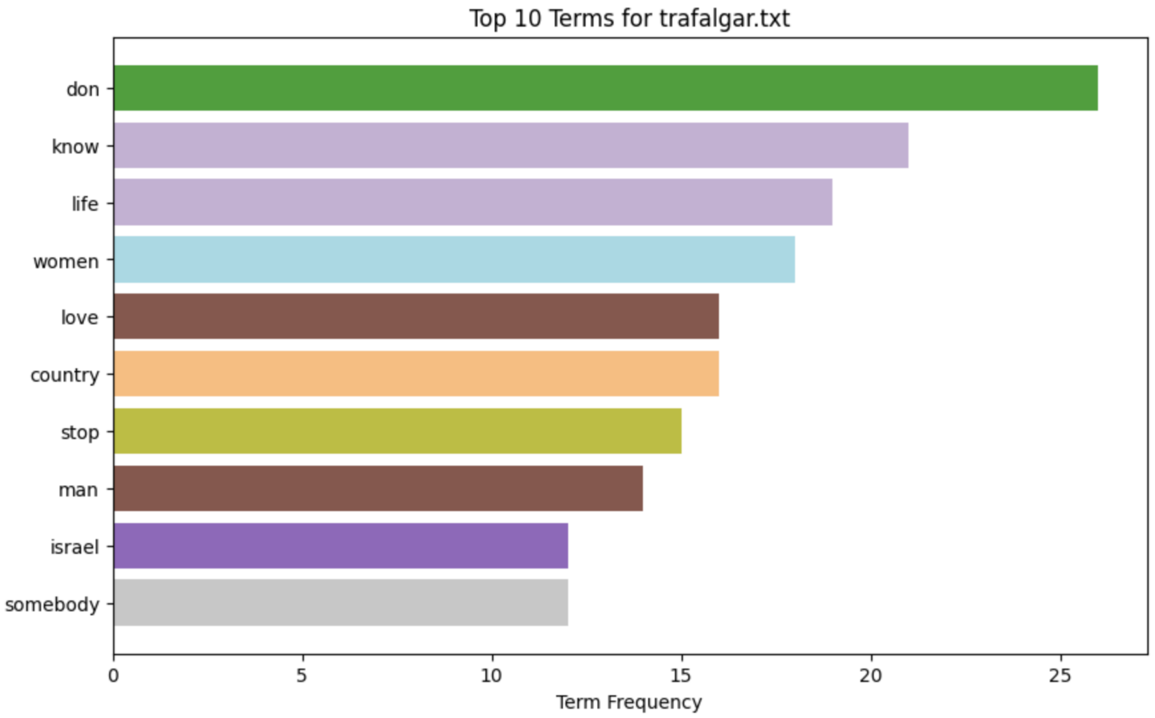

Figure 10: Top 10 Words in "Trafalgar"

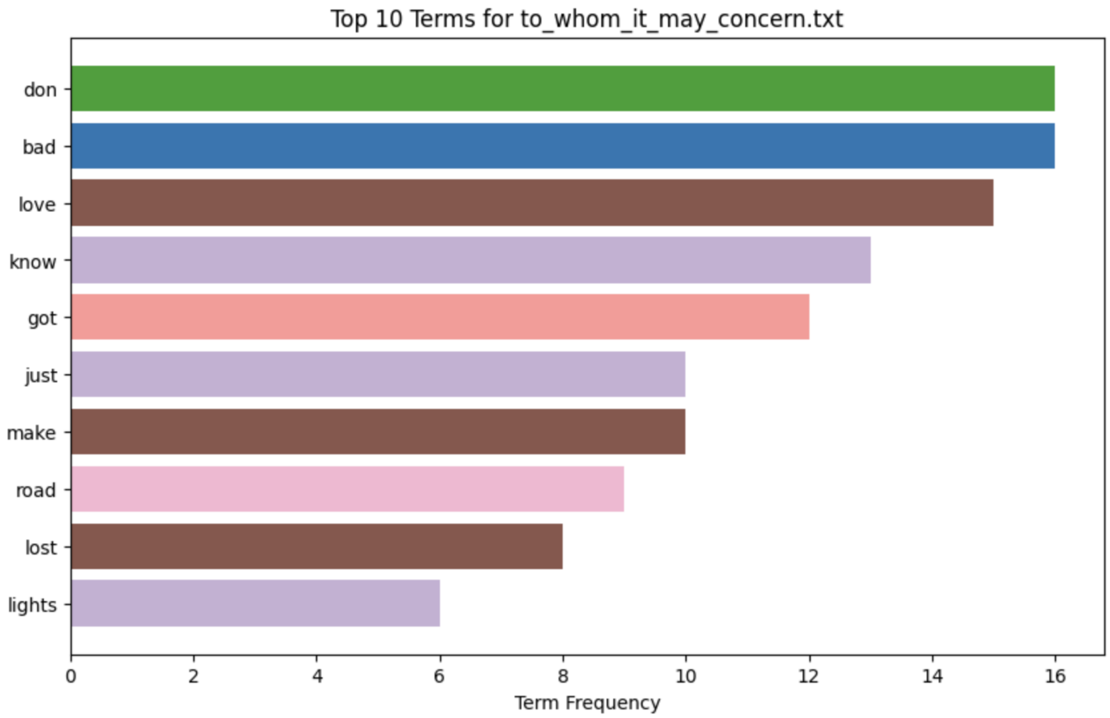

Figure 11: Top 10 Words in "To Whom It May Concern"

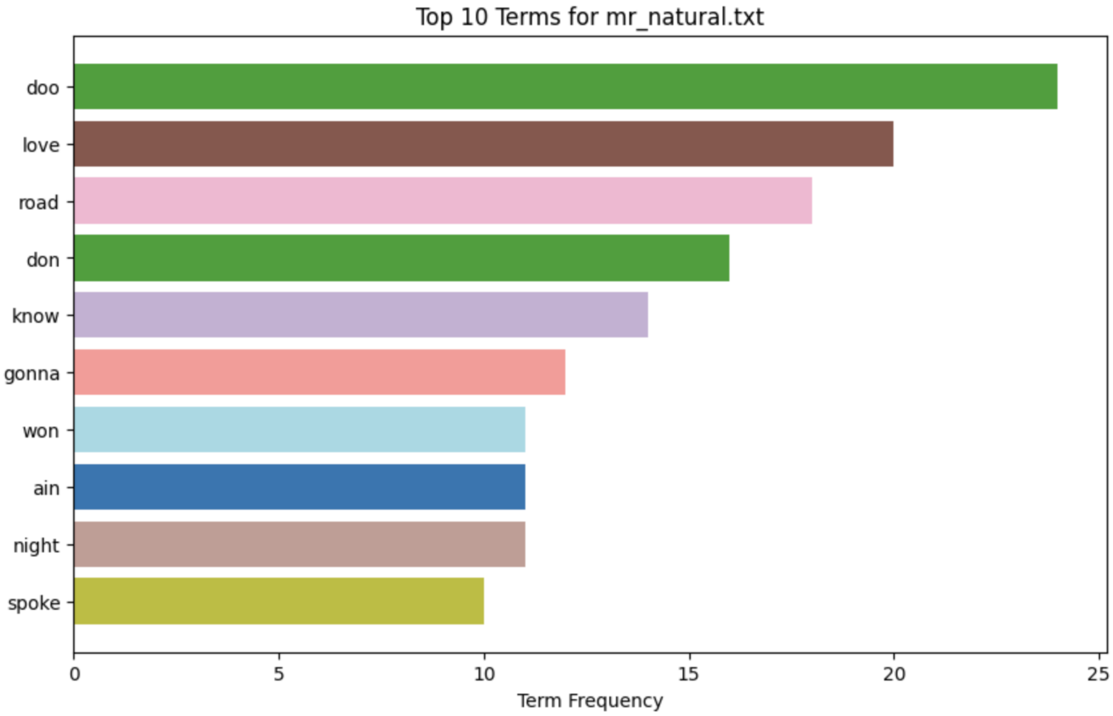

Figure 12: Top 10 Words in "Mr. Natural"

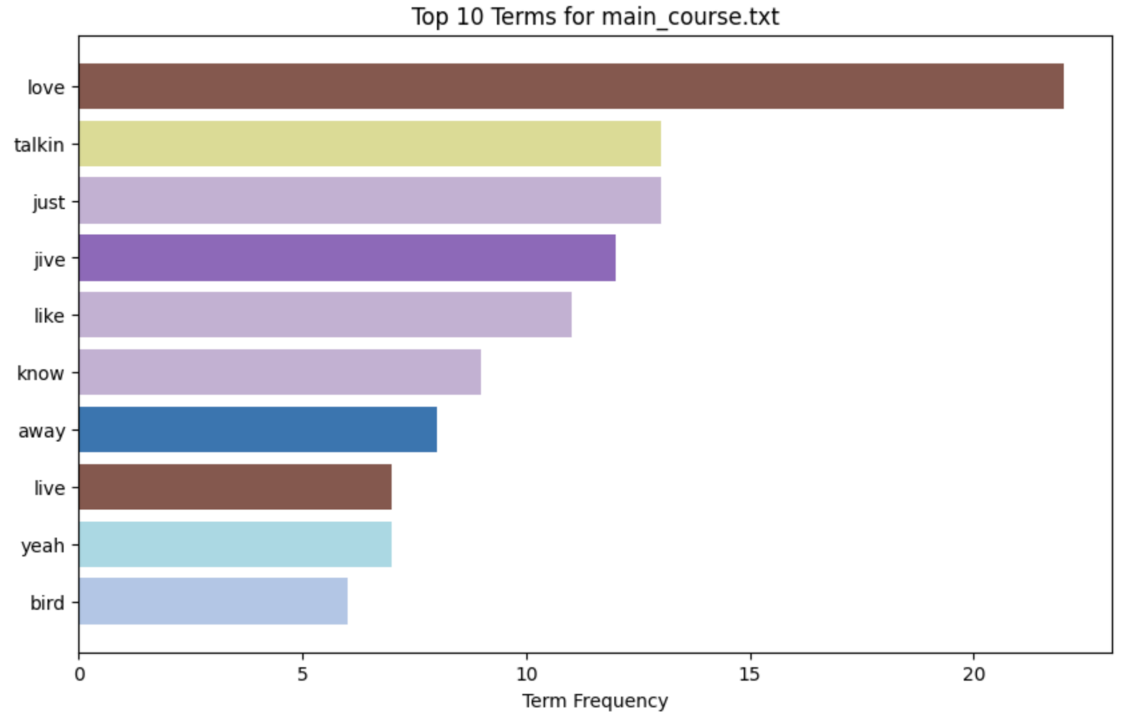

Figure 13: Top 10 Words in "Main Course"

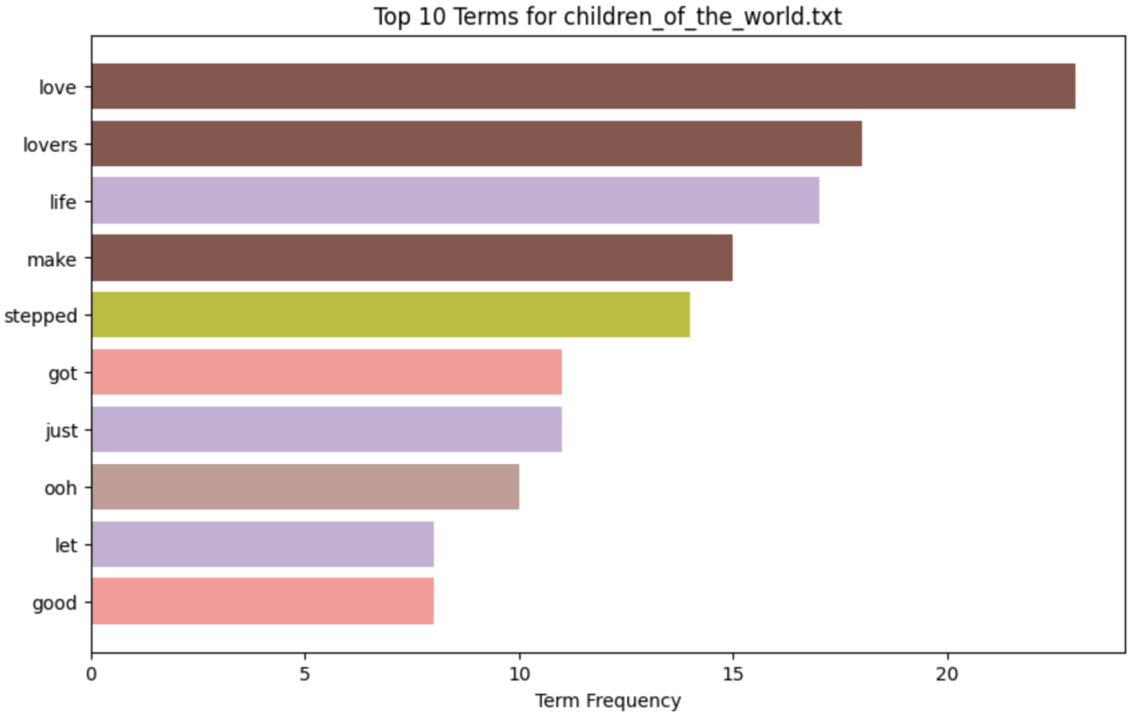

Figure 14: Top 10 Words in "Children of the World"

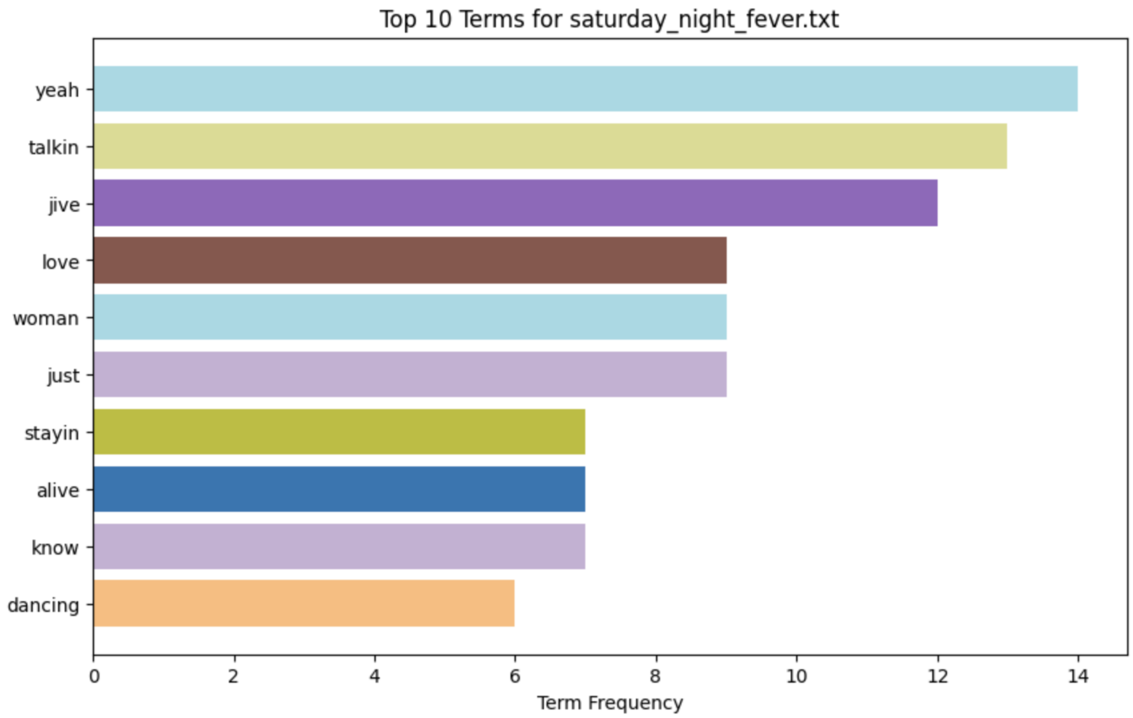

Figure 15: Top 10 Words in "Saturday Night Fever"

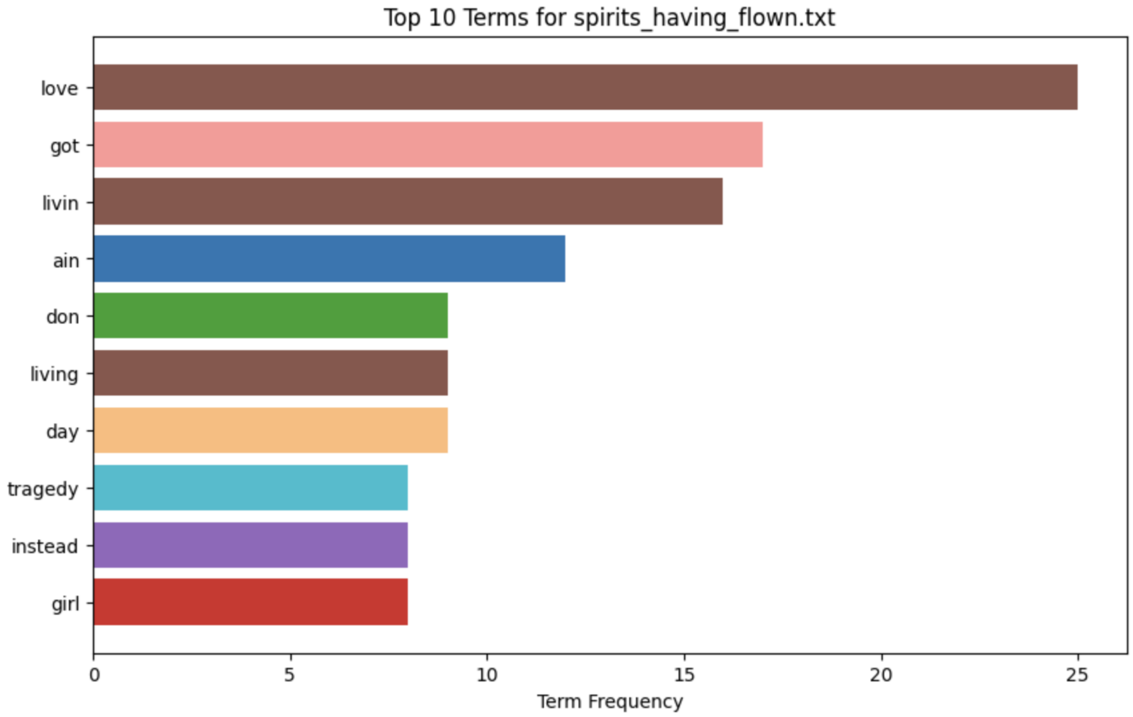

Figure 16: Top 10 Words in "Spirits Having Flown"

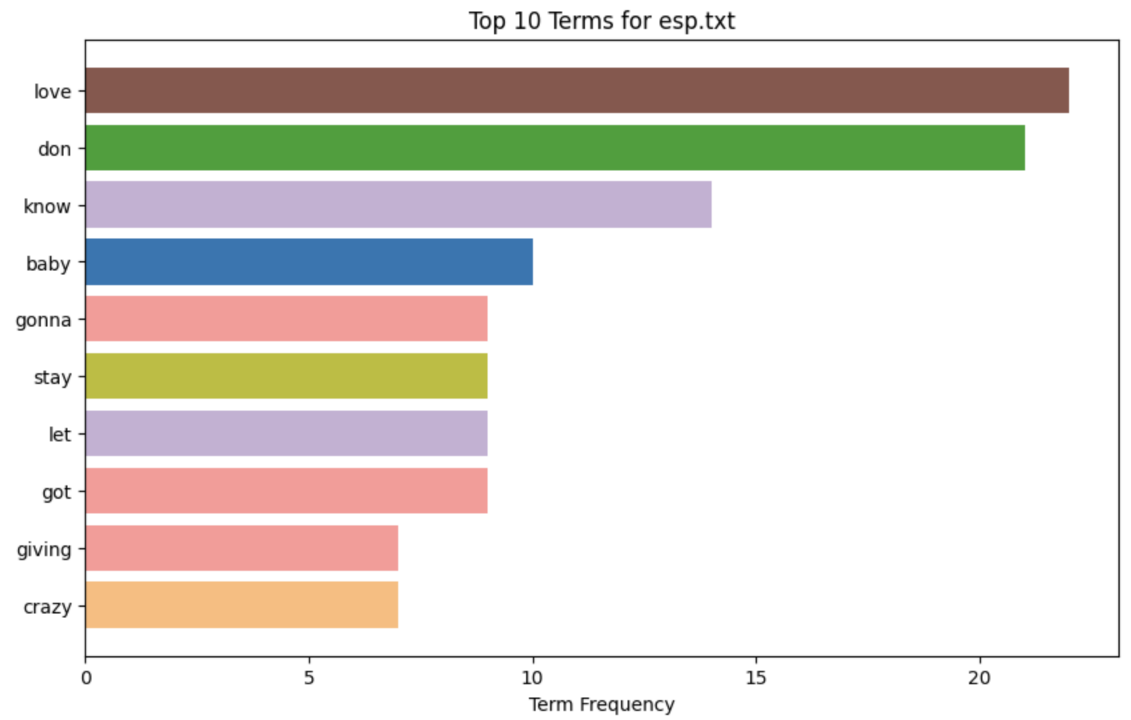

Figure 17: Top 10 Words in "E.S.P."

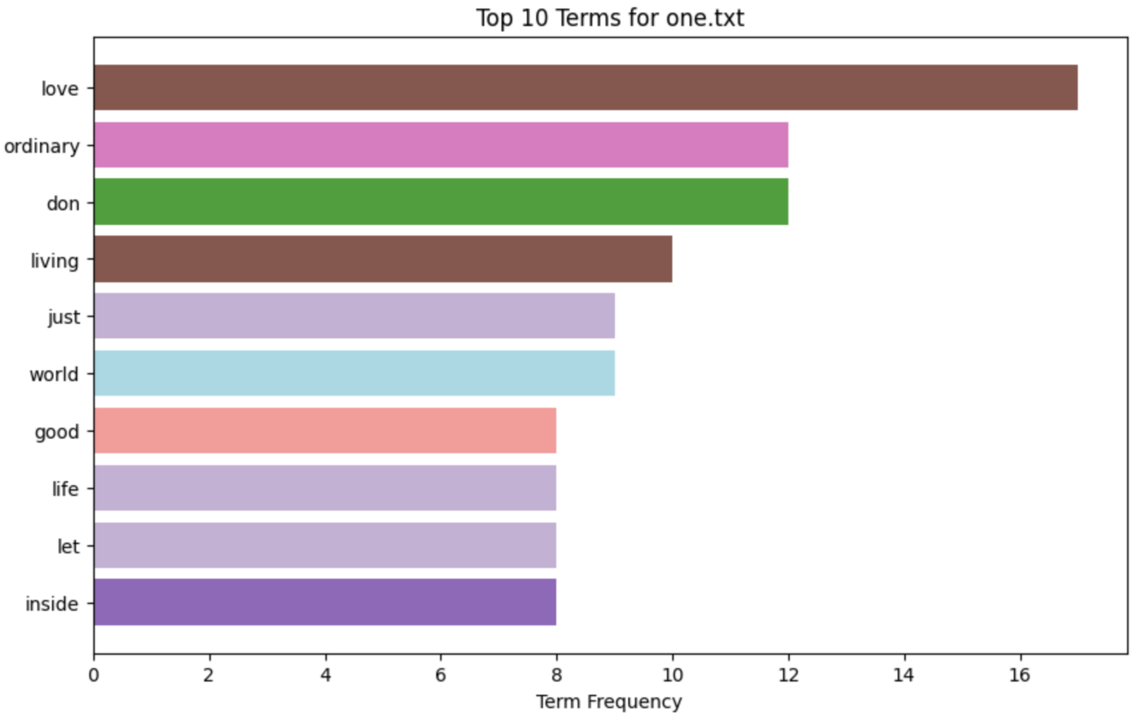

Figure 18: Top 10 Words in "One"

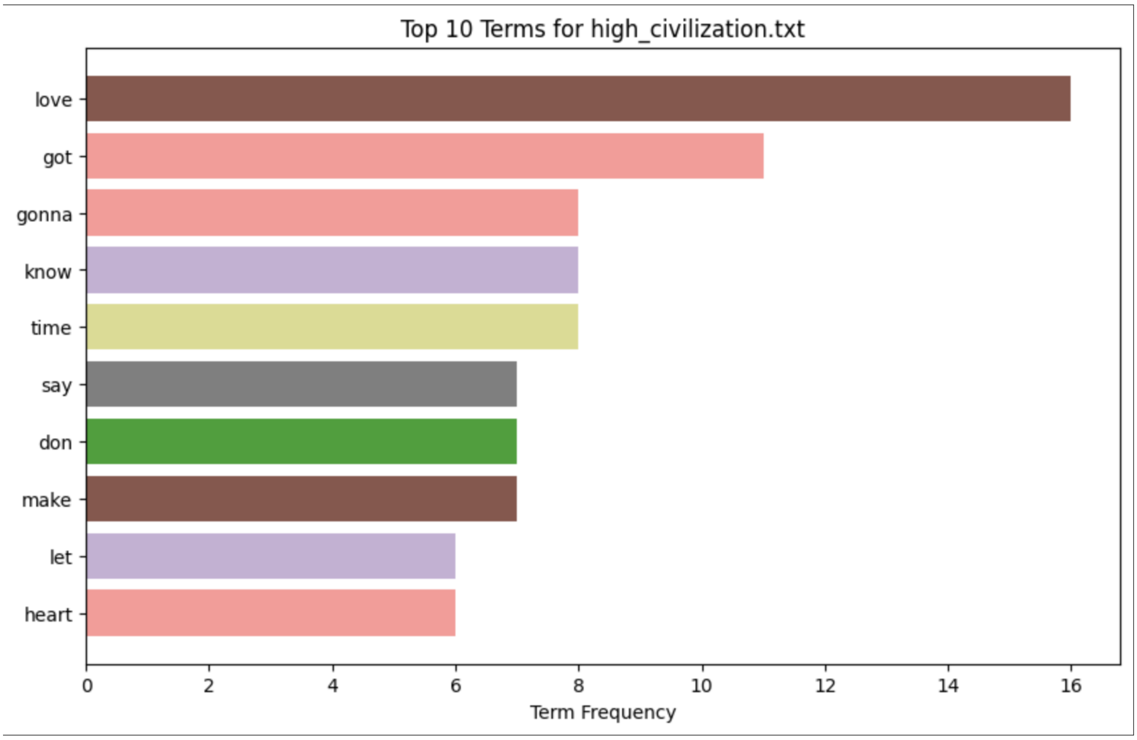

Figure 19: Top 10 Words in "High Civilization"

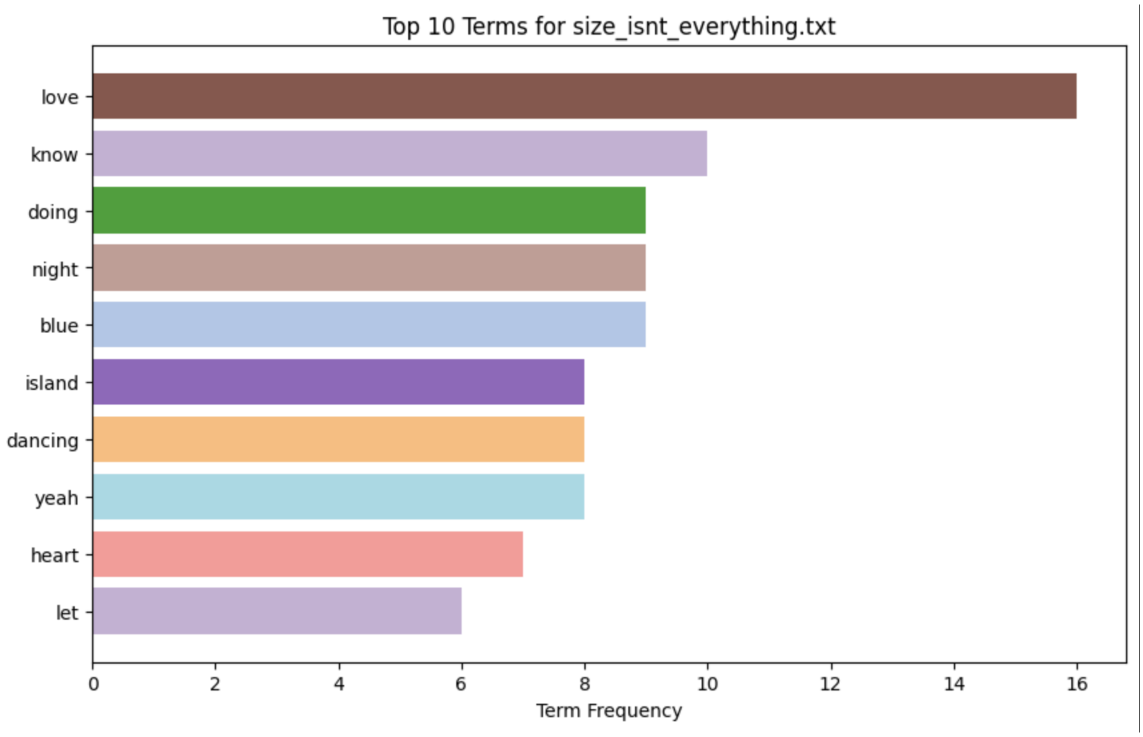

Figure 20: Top 10 Words in "Size Isn't Everything"

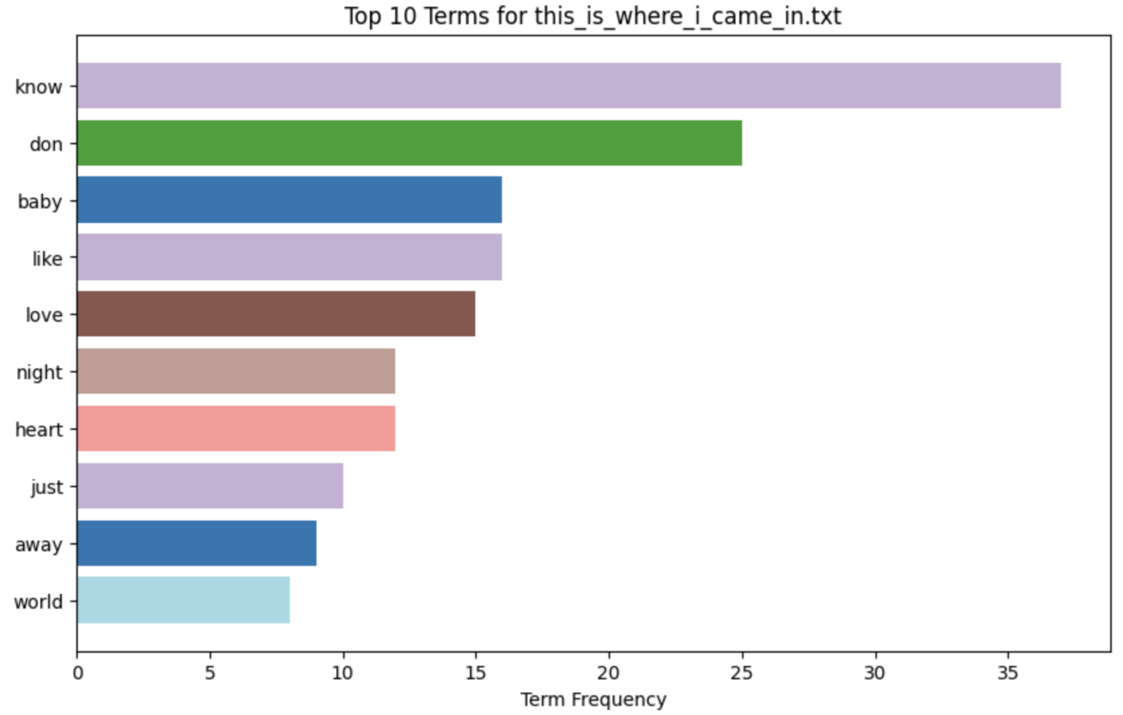

Figure 21: Top 10 Words in "This is Where I Came In"