Which Modern-Day Music Artist is William Shakespeare Most Similar to

Angel Sierra, Dominique McDonald, Mariana Gonzalez Castro,
Danny Ying, and Carina Kalaydjian

June 7th, 2022

Abstract

abstract text goes here here here here here here here here!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

1 Introduction

here we cite [4] our glorius leader [7] yeah yeah cool cool [10] and this too [9] this as well [14] woohooo!!! [13] and [6] and [8] and [1] sdljdj [3] , [12] and [11], [2], [5]

William Shakespeare is a world-renowned poet from 16th-century England. In his lifetime of 52 years, he wrote 37 plays, 154 sonnets, and many poems[1]. William Shakespeare has recognition for being one of the most revolutionary Literarry Artists to this date[1]. The Oxford Dictionary of Quotations states that William Shakespeare wrote close to one tenth of the most quoted lines ever written or spoken in English[1]. Shakespeare is also the second most quoted English writer after the writers of the bible[1]. Through his work, Shakespear introduced almost 3,000 words to the English language, these words can still be found in the Oxford English Dictionary today[1]! Shakespeare has an incredible ability to deeply express emotions through his work. Reders have been inspired and moved by his work for centuries. Shakespeare changed the world through his literary works. That is why William Shakespear is the best candidate to analyze and see how his Literary work compares to modern-day music artists song lyrics.

For this project we will be analzing Shakespeare's words and usage of emotion from his 154 sonnets from the Gutenberg Project[14]. We will then compare these words and emotions to another dataset

of 45 different modern-day music artists and their song lyrics[13]. Our interest is seeing which music artists song lyrics are the most similar to Shakespeare's sonnets in the aspect of two categories, similarities words used and similarities in emotions expressed. Then we will see which modern-day music artist is the most similar to Willam Shakspare.

In order to answer this question, we are going to use various methods of Text Analysis, analysis of unorganized text data[9]. The different forms of Text Analysis that we will be using for this project are Sentiment Analysis and Keyword Extraction. Sentiment Analysis involves analyzing the text and being able to extract the different emotions the text is expressing[9]. Keyword Extraction analyzes the text and identifies which words appear most frequently in the text, these words are called Keywords[9]. The combination of these two methods will bring the result of which music artist is the most similar to shakespear in therms of types of words used and which music artist is the most similar to Shakespeare in terms of emotion expressed through words.

2 Methods

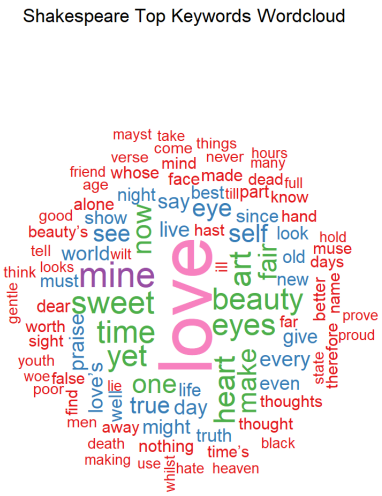
There is a growing demand for the application of Natural Language Processing to drive music knowledge discovery [11]. Within song lyrics there is a wealth of information that can be used to gain insights about the song and its listeners. The workflow used to conduct NLP is outlined in Oramas et. Al [12] as the following steps 1. Corpus creation (collection of separated documents), 2. Text mining (accessing desired info and eliminating the excess), 3. Information extraction (word frequency, collocation, word position, etc.), 4. Knowledge graph generation (a directed labeled graph in which we have associated domain specific meanings with nodes and edges [2]), 5. Sentiment Analysis (Identify feelings and emotions present in a text [9]).

This NLP pipeline can be applied to music to provide recommendations on what songs a listener might like based on what they already listen to. [11] This information is gathered by performing what's called a similarity search. Mahedero et al [11] conducts a similarity search for their project by first calculating a similarity measure. The similarity measure used is known as the Standard Cosine Distance (SCD). The calculations of the SCD are beyond the scope of this paper, but something to note is that the SCD relies on the Inverse Document Frequency as a way to measure the prevalence of words in a document and compare that to other documents.

As previously stated, we relied on Natural Language Processing methods to calculate the similarity between Shakespeare and current artists with the hope of identifying a single artist whose work has similar themes as Shakespeare's. Our pipeline was structured similarly to that proposed in [12] with the main difference being that we do not rely on knowledge graph generation to store any information or display findings. The methods that were most useful for this research were keyword extraction and sentiment analysis. Both methods of analysis provided us with crucial insights, so the details of each seem pertinent to share.

2.1 Keyword Extraction

Keyword extraction is largely aimed at identifying the most relevant words in different texts and utilizing those words to understand common theme or popular topics.[9] We used the tm library in R to perform keyword extraction[8] on the sonnets. This process involves first prepping the text by removing unwanted punctuation or numbers, eliminating stop words, changing everything to lower case. These transformations are necessary because when working with strings, you not only have to be precise, but you also must be exact. The cleaning process sometimes includes stemming the words, but we opted not to do that. Once the text is clean the idea is to create a table containing each word used in the text and the frequency with which it is used. To identify the top ten key words we sorted the words in the matrix by their frequency.



Once Shakespeare's keywords were identified, the next step was to identify the keywords for each of the 45 music artists. This process mirrored that of extracting keywords from the sonnets except we utilized an algorithm to automate the process for each artist. Once keywords were identified we converted the word frequencies for Shakespeare and all other artists into proportions, in order to standardize for comparison. To find the artists who were most like Shakespeare based on keywords, we checked for artist who had the highest number of matching keywords. Finally, then we ranked those by who used keywords in similar proportions to Shakespeare by calculating the Euclidian distance between the proportions of each of the matching keywords. This is similar to what Mahedero et. Al [11] does for their similarity search except we rely on word proportions and Euclidean distance, while their research utilizes Inverse Document Frequency and the Standard Cosine Distances.

Artist	Word Count	Frequency	Keyword	Word Rank
adele	3	626.00	love time heart	1
nickelback	3	11349.00	love time yet	2
bieber	2	648.00	love time	3
dolly-parton	2	685.00	love time	4
dj-khaled	2	928.00	mine time	5
amy-winehouse	2	3700.00	love time	6
bjork	2	5341.00	love yet	7
leonard-cohen	2	5972.00	love time	8
cake	2	7690.00	love time	9
paul-simon	2	8464.00	love time	10

Table 1: Ranked Top 10 Most Similar Music Artist to Shakespeare Based on Keywords

2.2 Sentiment Analysis

A similar sort of ranking was achieved from Sentiment Analysis, but the process has notable differences. The library syuzhet was used for the analysis. There is a wide variety of paths one could explore when performing sentiment analysis. Two that seemed viable for the purpose of this research were calculating an overall sentiment score and identifying the different emotions present in each text. [6] The latter option proved to be more fruitful, as the analysis is more detailed. This process of classification based on emotion is known as NRC sentiment analysis. NRC Sentiment Analysis uses the National Research Council (NRC) Word-Emotion Association Lexicon to classify words in a text into eight categories of emotions. [5] It is important to note that a word may be associated with more than one emotion. The eight emotions are anger, anticipation, disgust, fear, joy, sadness, surprise, trust (include simple equation). [5] The objective of NRC Sentiment Analysis is to calculate the frequency with which each emotion is conveyed. This is calculated by identifying the emotions associated with the unique words in a text and summing up all the instances of each emotion[6].

Artist	Sent. Euclidean Distance	Sentiment Rank
amy-winehouse	0.001227	1
eminem	0.001237	2
cake	0.001304	3
nirvana	0.001365	4
bob-dylan	0.001471	5
bob-marley	0.001492	6
johnny-cash	0.001589	7
nickelback	0.001827	8
britney-spears	0.002097	9
rihanna	0.002249	10

Table 2: Ranked Top 10 Most Similar Music Artist to Shakespeare Based on Sentiments

Once the NRC sentiments were calculated for both Shakespeare and the music artists, the frequencies were converted to proportions for accurate comparison. Using the proportions of each emotion, Euclidean distance was calculated for between each music artist and Shakespeare. The artists with the shortest distances from Shakespeare were considered the most similar to him, and therefore ranked higher in regard to comparison of the emotions conveyed in their works.

After comparing results from each analysis method, a final ranking was calculated. This overall ranking was calculated by summing up the rankings from both analysis methods. Because similarity to Shakespeare was assessed using Euclidean distance, lower rankings signify higher similarity to Shakespeare. This meant that the artists with the lowest overall ranking mirrored Shakespeare's work emotionally more than other artists.

Artist	Word Rank	Sentiment Rank	Overall Rank
nickelback	2	8	10
amy-winehouse	6	4	10
cake	9	3	12
adele	1	21	22
joni-mitchell	11	11	22
leonard-cohen	8	17	25
paul-simon	10	22	32
blink-182	18	14	32
bob-marley	27	6	33
rihanna	25	10	35

Table 3: Ranked Top 10 Most Similar Music Artist to Shakespeare

2.3 K-Means Clustering

The NRC sentiments were utilized even further as predictors in K-Means Cluster Analysis. The objective of cluster analysis within the scope of this research is to utilize an unsupervised learning model [3] to assess commonalities between the work of each music artist and Shakespeare. K-Means clustering is an iterative method that categorizes each data point into one of k predefined groups, or clusters. The process is driven by two objectives. The first being maximizing the distance between clusters, so that they are distinct. The second is minimizing the data points within a cluster, so the clusters themselves are homogenous. [3] By employing K-means Clustering we were able to identify a group of artists that whose work most closely matches Shakespeare's.

2.4 Data Management

Before discussing the results of our various methods, it is important to address what made such an undertaking possible: data management. The size of this project necessitated multiple team members using multiple platforms. The data was stored in a relational database and then hosted on Amazon Web Services(AWS) service called Relational Database Services(RDS). Using RDS preserves the database from alterations, each memeber each member provided the read only user credentials to their database to enable the team to access data without making changes to the database itself. Once we had the data it the next steps were to process and analyze it.

Again, with so many contributors working to advance the project it was necessary to have an avenue for efficient and organized sharing of code. For this aspect of the project, GitHub was

utilized and it allowed team members to work on the same files from different locations and share them as frequently as necessary. Along with our code we are also able to store and share important information that aided us in our research. The different information sharing structures employed allowed for efficient progression and ultimately valuable results.

3 Results

4 Conclusions

References

[1] 56 Fun Facts about William Shakespear. https://nosweatshakespeare.com/resources/shakespeare-facts/. Accessed May 31, 2022.

[2] An Introduction to Knowledge Graphs. http://ai.stanford.edu/blog/introduction-to-knowledge-graphs/. Accessed June 3, 2022.

[3] K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. Accessed May 31, 2022.

[4] Natural Language Processing With R. https://www.udacity.com/blog/2020/10/natural-language-processing-with-r.html. Accessed May 20, 2022.

[5] NRC Word-Emotion Association Lexicon. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. Accessed June 3, 2022.

[6] Syuzhet Package Documentation in R. https://www.rdocumentation.org/packages/syuzhet/versions/1.0.6. Accessed May 20, 2022.

[7] Text Mining and Sentiment Analysis: Analysis with R. https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-with-r/. Accessed May 20, 2022.

[8] Tm: Text Mining Package Documentation in R. https://rdrr.io/rforge/tm/man/. Accessed May 20, 2022.

[9] What is Text Analysis? A Beginner's Guide. https://monkeylearn.com/text-analysis/. Accessed May 20, 2022.

[10] Kristin Briney. *Data Management for Researchers: Organize, maintain and share your data for research success*. Pelagic Publishing Ltd, 2015.

[11] Jose P. G. Mahedero, Álvaro MartÍnez, Pedro Cano, Markus Koppenberger, and Fabien Gouyon. Natural language processing of lyrics. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, page 475–478, New York, NY, USA, 2005. Association for Computing Machinery.

[12] Sergio Oramas, Luis Espinosa-Anke, Francisco Gómez, and Xavier Serra. Natural language processing for music knowledge discovery. *Journal of New Music Research*, 47(4):365–382, 2018.

[13] Paul Mooney. Song Lyrics: Poetry and Lyric (TXT files). Accessed May 20, 2022.

[14] William Shakespear. Shakespear's Sonnets. Project Gutenberg, 1997 [Online]. EBook-No. 1041, Accessed May 20, 2022.