The Story of National Anthem Through Machine Learning

Chakraborty, Deepankar [1]

[1]The City College of New York, CSc 44700 Machine Learning

1. Abstract

In this study, we used natural language processing (NLP) techniques and k-means clustering to cluster a set of 190 national anthems based on their English lyrics. National anthems play a significant role in shaping a country's national identity and promoting unity and pride among its citizens. They are often used to express national values and ideals. In this paper, we will identify the patterns and themes that can be shared by groups of countries with similar cultural or political histories by clustering them into different categories. We first preprocess the lyrics of the national anthems by removing stop words and stemming the remaining words. We then use NLP techniques called Tf-Idf, which can help identify the most important words in a document or a collection of documents. Then We then used the k-means algorithm to cluster the anthems into five groups based on the similarity of their lyrics using the Tf-Idf vector of numbers. Then we tested and evaluated the performance of the model using exploratory data analysis method. Our results showed that the model was able to quite accurately group the national anthems into clusters that reflected shared themes and cultural influences. This study provides a novel approach to analyzing national anthems and has the potential to inform research on cultural and political trends, as well as to serve as a resource for educational applications for understanding what role does anthem play in a nation's identity.

Keywords: *Machine Learning, Natural Language Processing, Tf-idf, k-means clustering, BERT*

Table of Contents:

2. Introduction

A national anthem is a song or music that is associated with a particular country and is intended to evoke national pride and unity. It is typically played or sung at official ceremonies, such as during the opening or closing of a national sports event, or on national holidays. National anthems can serve several important purposes. For example, it can express national identity which makes the said country stand-out among others. National anthems are a way for a country to express its unique identity and cultural heritage. They often include references to the country's history, geography, or values, and can help to foster a sense of pride and unity among its citizens.

3. Background and Motivation

Clustering national anthems into different categories based on their lyrics can provide insight into the themes and values that are important to a particular country. For example, some national anthems may focus on themes related to nature, such as the beauty of the country or the bravery of its people. Others may focus on themes related to the nation's history or cultural traditions. Still, others may focus on themes related to love and devotion to the nation or its leader.

The motivation for this project was lyrical analysis of the national anthems, by clustering them, to identify patterns and themes that may be shared by groups of countries with similar cultural or political histories. This could give a new lens of how we perceive a national anthem reveals the factors that shape national identity from the anthem and can help gain a deeper understanding of the values and themes that are important to a particular country.

4. Algorithms:

Clustering national anthems using lyrics with machine learning involves using several algorithms and techniques like Natural Language Processing(NLP), k-means clustering, etc. The outline of the steps is described below:

1. Gather the lyrics for a set of national anthems that needs to be clustered. A pre-compiled dataset of 190 national anthems in English lyrics has been collected from Kaggle.
2. Preprocess the lyrics by cleaning and normalizing the text. This includes removing punctuation, lowercasing all words, and stemming or lemmatizing the words to reduce them to their base forms, removing nation's names, and nationality names.
3. Represent the lyrics as numerical data that can be used as input to a machine learning algorithm. The approach used in this paper is called term frequency-inverse document frequency (TF-IDF) vectors, which measure the importance of each word in the lyrics relative to the entire set of lyrics.
4. Choose a clustering algorithm to use. There are many different algorithms available, including k-means, hierarchical clustering, and density-based clustering. For this paper, the k-means clustering algorithm is used for categorizing the anthems.
5. Train the chosen algorithm on the preprocessed lyrics data. This will involve specifying the number of clusters to generate and any other relevant parameters.
6. Visualize the clusters using a technique like principal component analysis (PCA) or 2D mapping, and exploratory data analysis of the data and plot it in two dimensions. This helps with understanding the relationships between the national anthems and seeing the relationships within a group.

4.1 Feature Extraction with TF-IDF

TF-IDF (term frequency-inverse document frequency) is a common technique used in natural language processing (NLP) to measure the importance of a word in a document or a collection of documents. It is based on the idea that a word that frequently occurs in a document is more important than a word that occurs less frequently, but that the importance of a word also depends on how common it is across all documents.

Mathematically, the TF-IDF score of a word w in document d is calculated as follows:

Term Frequency(TF):

$$TF(w,d) = \frac{Number\ of\ occurrences\ of\ w\ in\ d}{Total\ number\ of\ words\ in\ d}$$

Inverse Document Frequency (IDF):

$$IDF(w) = \log(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ w})$$

The TF part of the equation measures the frequency of a word in a single document, while the IDF part measures the rarity of the word across all documents. The combination of these two measures gives a score that reflects the importance of the word in the given document.

Overall, Tf-Idf:

$$TF-IDF(w,d) = TF(w,d) \times IDF(w)$$

For instance, if a word occurs frequently in a document but is also common across all documents, its TF-IDF score will be low because the IDF part of the equation will be small. On the other hand, if a word occurs frequently in a document and is rare across all documents, its TF-IDF score will be high because the IDF part of the equation will be large.

4.2 K-means clustering:

K-means clustering is an unsupervised machine learning algorithm that is used to partition a dataset into a predefined number of clusters (also known as "k"). The goal of the algorithm is to minimize the within-cluster variance, or the sum of the squared distance between each data point and the centroid (i.e., the mean of the data points) of its cluster.

To begin the algorithm, k initial centroids are randomly chosen from the data points. The algorithm then iteratively performs the following two steps until convergence:

a. **Assignment step**: Each data point is assigned to the cluster with the nearest centroid. This can be done by calculating the Euclidean distance between the data point and each centroid and choosing the centroid with the smallest distance. Let's say we have a dataset of n data points, represented by the points x1, x2, ..., xn, and k centroids, represented by the points µ1, µ2, ..., µk. We can represent the assignment of each data point to a cluster using a set of binary variables y1, y2, ..., yn, where yi = 1 if xi is assigned to cluster j and yi = 0 otherwise. The objective of the k-means algorithm is to minimize the within-cluster sum of squares (WCSS), which is defined as:

$$\mathrm{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \text{cluster } j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2,$$

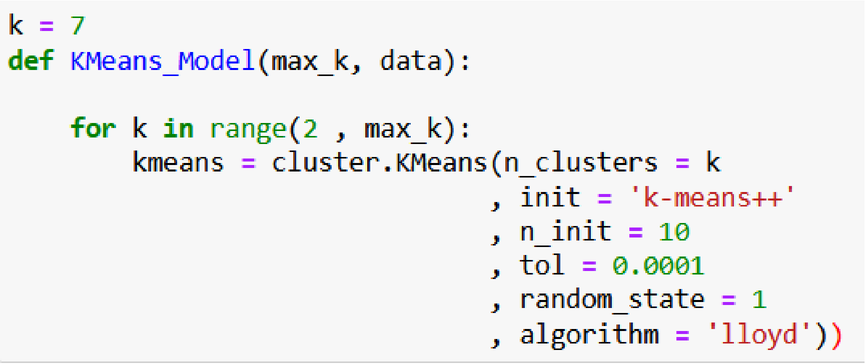
where $\bar{\mathbf{x}}_j$ is the sample mean in cluster j

b. **Update step**: The k-means algorithm tries to find the values of the centroids μ1, μ2, ..., μk that minimize the WCSS. The algorithm iteratively improves the centroids by repeatedly reassigning the data points to the closest centroid and then recomputing the centroids based on the mean of the data points in each cluster. So, the centroids are recalculated for each cluster as the mean of all the data points in the cluster. The algorithm stops when the centroids no longer change or a maximum number of iterations is reached.

One of the main advantages of the k-means clustering is its simplicity and efficiency. It is a fast algorithm that can handle large datasets, and it is relatively easy to implement and understand. However, it is sensitive to the initial centroid selections and can sometimes produce suboptimal results, particularly if the clusters are not spherical or if the data is noisy or unevenly distributed.

5. Model:

Scikit-learn is a popular Python library that provides tools for machine learning and data analysis. It includes an implementation of the k-means clustering algorithm that can be used to group data points into a specified number of clusters.



```
k = 7
def KMeans_Model(max_k, data):

    for k in range(2 , max_k):
        kmeans = cluster.KMeans(n_clusters = k
                                , init = 'k-means++'
                                , n_init = 10
                                , tol = 0.0001
                                , random_state = 1
                                , algorithm = 'lloyd'))
```

Figure: K-means cluster model initialization

- For creating the model, the first step is to get the cleaned and transformed data. The preprocessed data is saved in a data frame ready to be used by our model.
- Then **Import the k-means model:** To use the k-means model in scikit-learn, it needs to be imported from the `` `cluster` `` module.
- **Instantiate the model:** An instance of the k-means model needs to be created by calling the KMeans class and specifying the number of clusters (k). Initially, the number of clusters chosen was, `n_cluster=7.`
    - Then the tolerance(tol) was specified. Tol is the "relative tolerance with regards to Frobenius norm of the difference in the cluster centers of two consecutive iterations to declare convergence." Essentially it is a distance measure that will be used to assign points to clusters (e.g., Euclidean distance).
    - Then the method for initializing the centroids was chosen. Random sampling or `k-means++` both are valid options for this purpose. For this model, kmeans++ was chosen, which "selects initial cluster centroids using sampling based on an empirical probability distribution of the points' contribution to the overall inertia."(Sklearn-sklearn.cluster.KMeans.html)
- **Fit the model to the data**: Once the model is instantiated, the model can fit to the data by calling the `fit` method and passing in the data as an argument. This will train the model and calculate the centroids for each cluster.

5.1 Elbow Method

Elbow for finding the right cluster count.

The elbow method is a heuristic used in determining the optimal number of clusters in a k-means clustering algorithm. It is done by plotting the within-cluster sum of squares (WCSS) for each possible value of k and selecting the elbow of the plot as the optimal number of clusters.

To use the elbow method, we first need to compute the WCSS (within-cluster sum of squares) for a range of values of k. We can start with a small number such as 1 and increase to a larger number, such as 10. Then WCSS can be computed by fitting the k-means clustering algorithm to the data for each value of k and then calculating the WCSS.

Mathematically, the WCSS for a given value of k can be expressed as:

$WCSS(k) = \sum (x - \mu)^2$ **,** where $\boldsymbol{x}$ is a data point, $\boldsymbol{\mu}$ is the centroid of the cluster to which x belongs, and the sum is taken over all data points in the dataset.

Sklearn provides a handy method called ***`kmeans.inertia_`*** which returns the sum of squared distances of samples to their closest (centroid) cluster center. After getting all the WCSS values we can plot them in the following graph:

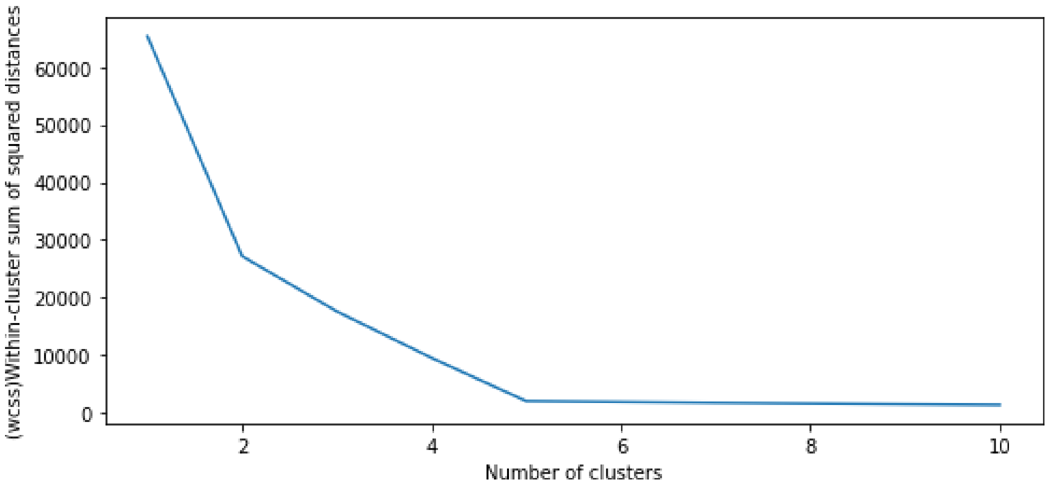


Figure: The Elbow chart- x-axis shows a number of clusters, and the y-axis shows WCSS score.

The above graph shows the relation between the number of clusters on the x-axis, and WCSS(sum of squared distances) on the y-axis. Then we need to find the 'elbow' point which anatomically looks like a human elbow. The Elbow point is a point where WCSS begins to decrease more slowly. In our project, the optimal value of k is 5, as indicated by the elbow in the plot above. This means that the data can be most accurately partitioned into five clusters, which are well-separated and compact. Beyond 5, we get a diminishing marginal return, where increasing the number of clusters beyond this point does not significantly improve the WCSS or the clustering outcome.

It is also worth noting that the elbow method is a heuristic and does not guarantee the optimal number of clusters. It may be necessary to choose a different final value for k and use other methods, such as visual inspection of the clusters or external validation measures, to determine the best number of clusters for a given dataset.

6. Result

The result of the model can be visualized below:

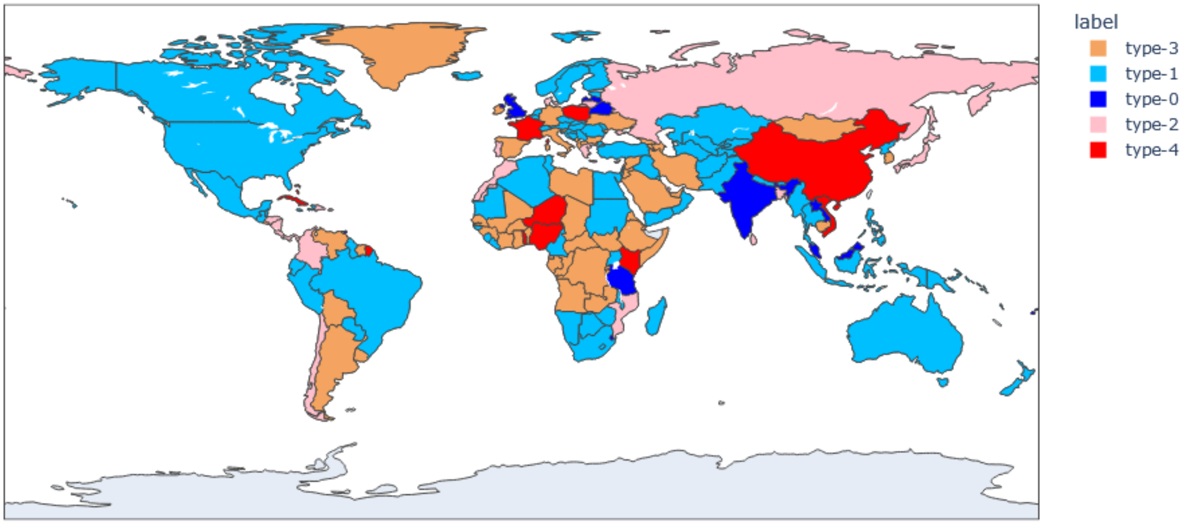


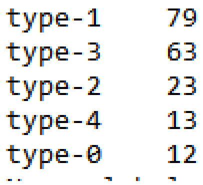
Figure: Cleropleth Map showing the national anthem cluster map

After selecting the optimal number of k, a new column on the dataset was added that indicated the cluster that country belongs to.

- Cluster 0: Blue
- Cluster 1: Sky blue
- Cluster 2: Pink
- Cluster 3: Brown
- Cluster 4: Red

From the map it can be seen that there aren't any statistically significant influences of geography in the clustering of the national anthems. All types of national anthems are randomly and uniformly distributed across 6 different continents. Cluster-0 primarily consists of countries from Asia and Europe, and3not include the Americas or other continents.

The distribution of the 5 groups are below:



```
type-1    79
type-3    63
type-2    23
type-4    13
type-0    12
```

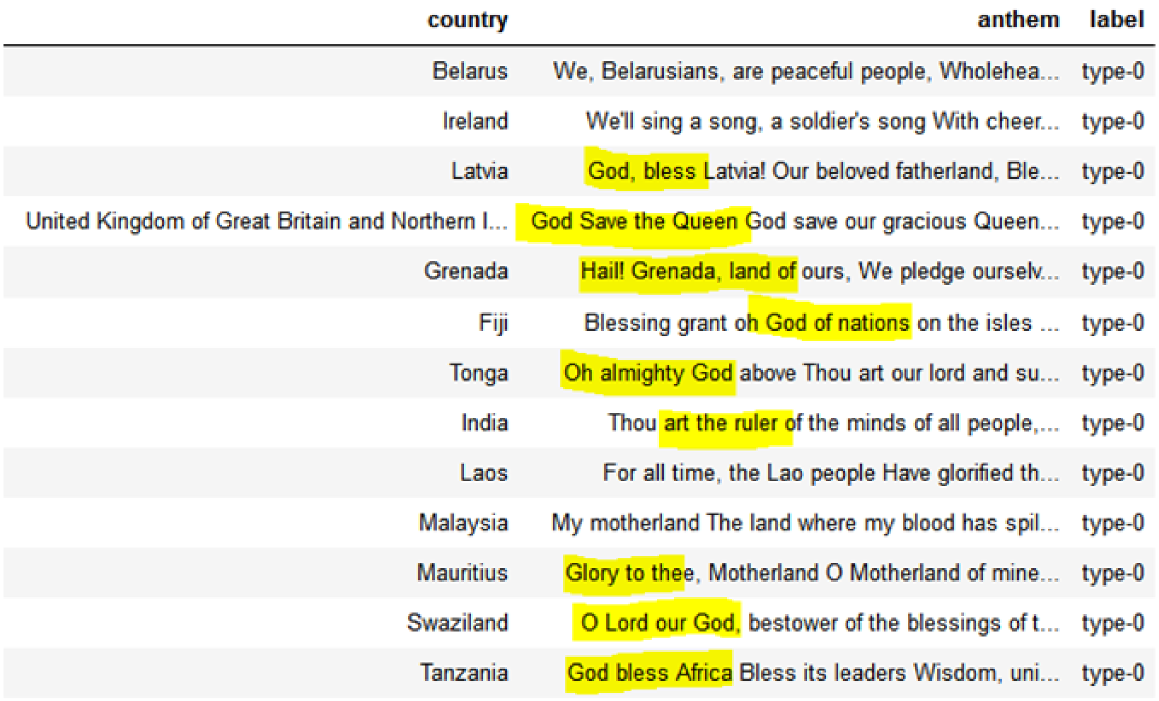
It can be seen that cluster-0 has the least number of anthems and whereas cluster-1 has the most.

7. Discussion

From the result section, it can be seen that national anthems can be clustered into different categories based on the content and themes of their lyrics. The analysis of the different clusters are discussed below:

Cluster-0 - (Praise the God and the King/Queen Cluster):

The countries(limited numbers) in this cluster are as follows:



country	anthem	label
Belarus	We, Belarusians, are peaceful people, Wholehea...	type-0
Ireland	We'll sing a song, a soldier's song With cheer...	type-0
Latvia	God, bless Latvia! Our beloved fatherland, Ble...	type-0
United Kingdom of Great Britain and Northern I...	God Save the Queen God save our gracious Queen...	type-0
Grenada	Hail! Grenada, land of ours, We pledge ourselv...	type-0
Fiji	Blessing grant oh God of nations on the isles ...	type-0
Tonga	Oh almighty God above Thou art our lord and su...	type-0
India	Thou art the ruler of the minds of all people,...	type-0
Laos	For all time, the Lao people Have glorified th...	type-0
Malaysia	My motherland The land where my blood has spil...	type-0
Mauritius	Glory to thee, Motherland O Motherland of mine...	type-0
Swaziland	O Lord our God, bestower of the blessings of t...	type-0
Tanzania	God bless Africa Bless its leaders Wisdom, uni...	type-0

Many national anthems include themes of praise for a god or a monarch or leader. This can take the form of expressing loyalty and devotion, thanking the deity or leader for their protection and guidance or simply glorifying their greatness and power. I labeled this cluster as the ‘Praise the God and the King/Queen’ cluster. From the lyrics highlighted in yellow, it can be seen that the major theme across the anthems in this cluster is dedicated towards praying for their homeland to God or praising their King/Queen.

For example, the national anthem of the United Kingdom, "God Save the Queen," begins with the lyrics "God save our gracious Queen, / Long live our noble Queen, / God save the

Queen!" and goes on to ask God to "send her victorious, / Happy and glorious, / Long to reign over us."

Similarly, the national anthem of Saudi Arabia, "العلي العظيم" (translated as "The Greatness of the Most High"), includes the lyrics "O God, our Lord, the Greatness of the Most High / Bestow upon our King the glory and the victory".

In both of these examples, the national anthem is expressing praise and devotion to a higher power or leader and calls on that power for protection and guidance.

The word cloud for this cluster can be seen in Appendix 1.

Cluster- 1- (A mix of different themes)

Cluster 1 has the most number of countries with 79 counts.

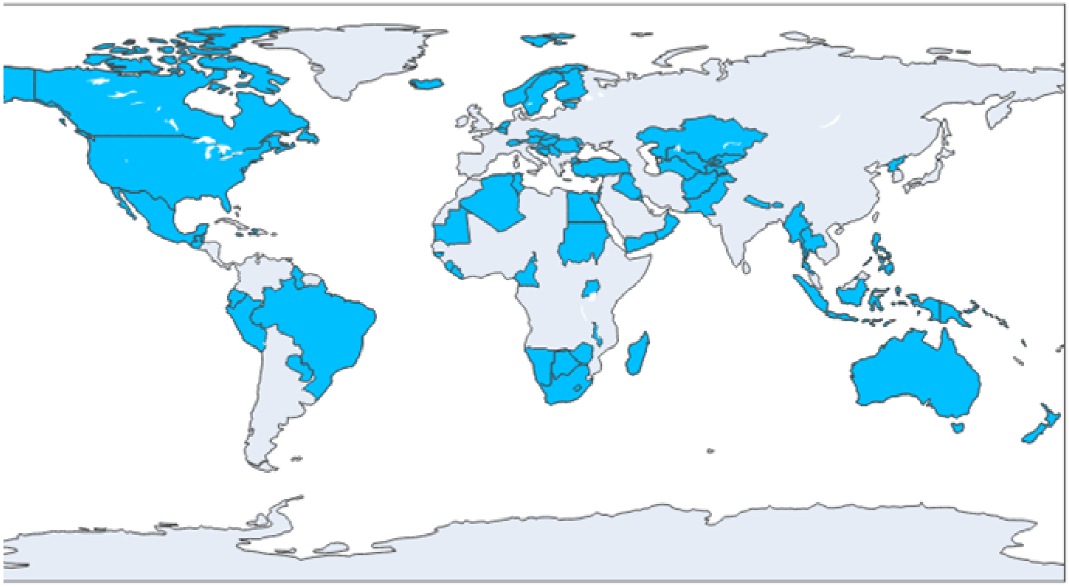


Figure: Cleropleth Map showing the national anthem cluster of only cluster-1

This cluster had the most variation in the lyrics. They don't have definitive themes, instead they present a wide range of themes including general patriotism, nature, God, praying etc.

Cluster- 2 - (Motherly and Nature Cluster)

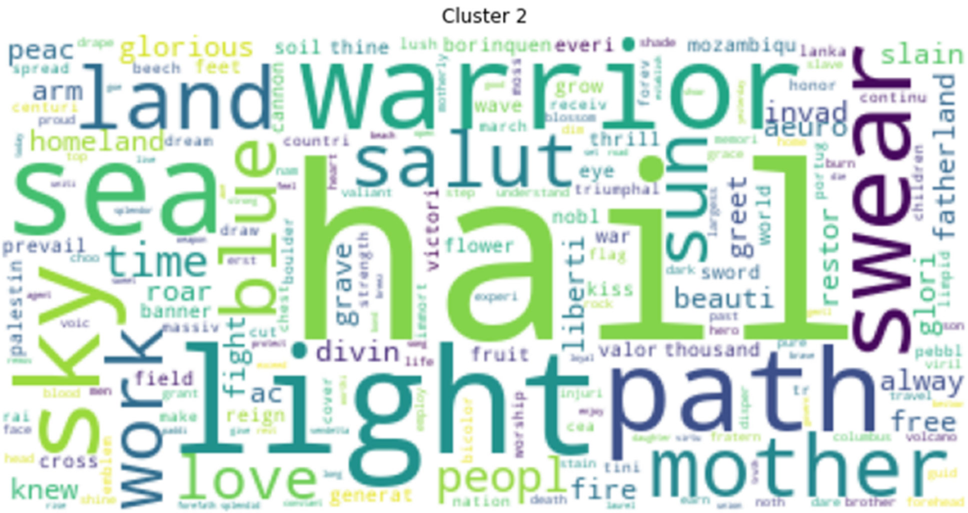
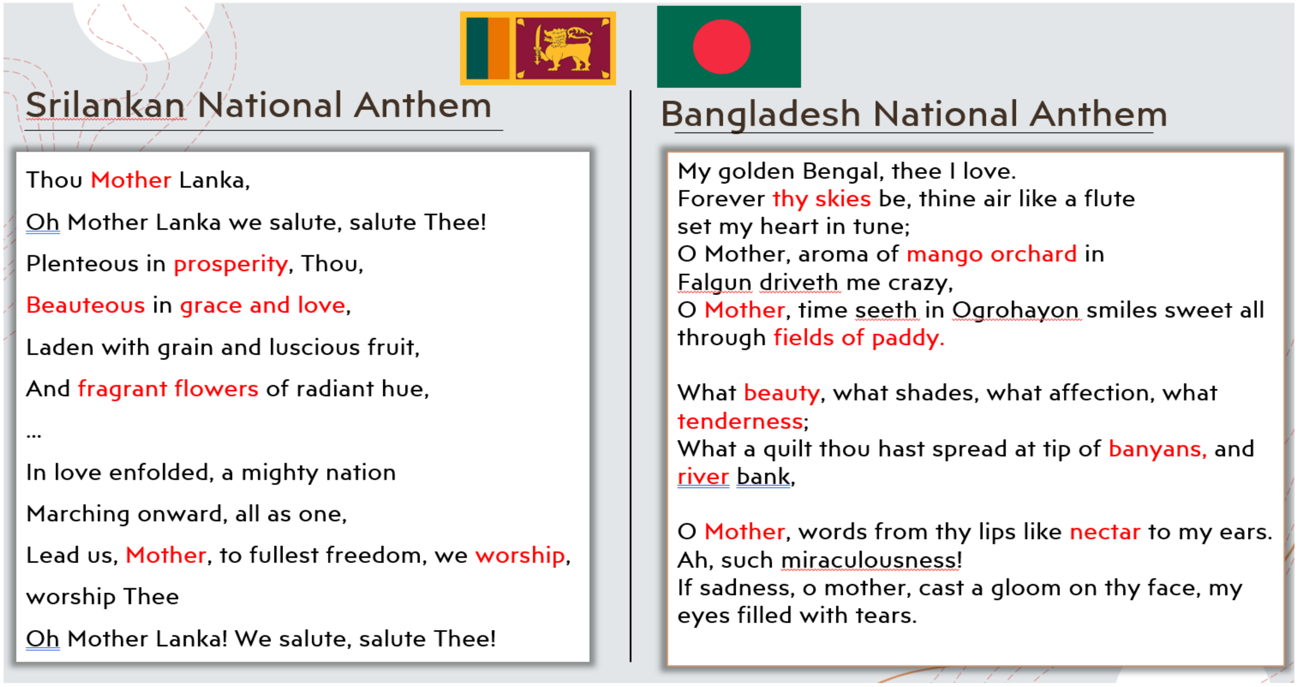

Cluster 2

Figure: Word Cloud showing the most important words in cluster 2.

Countries in this cluster include Bangladesh, Srilanka, Russia, Mozambique, Dominican Republic, Western Sahara, etc. Let’s now compare the lyrics of Bangladesh and Srilanka, which are part of this cluster:



Srilankan National Anthem

Thou Mother Lanka,

Oh Mother Lanka we salute, salute Thee!

Plenteous in prosperity, Thou,

Beauteous in grace and love,

Laden with grain and luscious fruit,

And fragrant flowers of radiant hue,

...

In love enfolded, a mighty nation

Marching onward, all as one,

Lead us, Mother, to fullest freedom, we worship, worship Thee

Oh Mother Lanka! We salute, salute Thee!

Bangladesh National Anthem

My golden Bengal, thee I love.
Forever thy skies be, thine air like a flute
set my heart in tune;
O Mother, aroma of mango orchard in
Falgun driveth me crazy,
O Mother, time seeth in Ogrohayon smiles sweet all through fields of paddy.

What beauty, what shades, what affection, what tenderness;
What a quilt thou hast spread at tip of banyans, and river bank,

O Mother, words from thy lips like nectar to my ears.
Ah, such miraculousness!
If sadness, o mother, cast a gloom on thy face, my eyes filled with tears.

From the above comparison, it can be seen that Srilanka and Bangladesh both have elements of nature and mother in their national anthem, that's why they were grouped together. Additionally, if we look at a part of the Russian national anthem, we can also find it revolves around the theme of the beauty of their nation:

"From the southern seas to the polar edge/ Our forests and fields are spread out.
You are the only one in the world! You are the only one –
the native land so kept by God!"

In this cluster of national anthems, nature is depicted as a source of beauty and strength, and the motherly figure is often seen as a symbol of protection and nurture, that's why I called this cluster as Motherly love and Nature cluster.

Cluster-3 (Freedom and Patriotic)

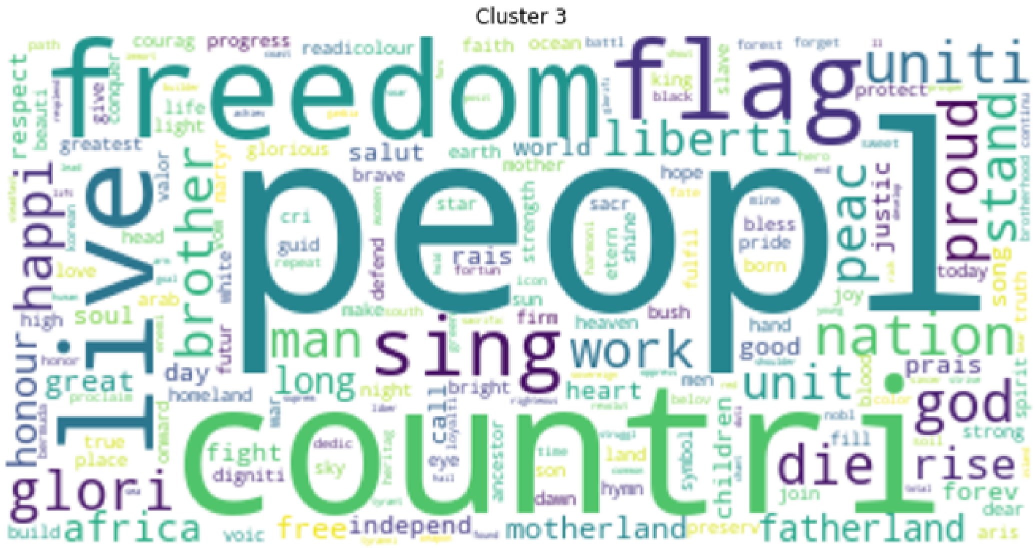
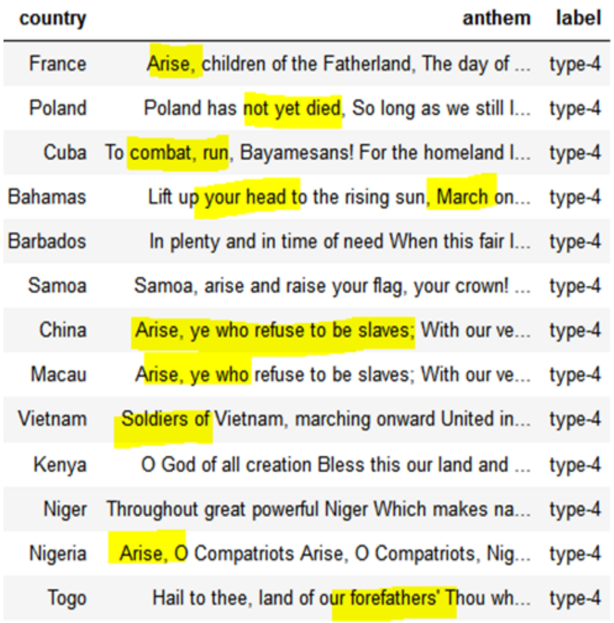


Figure: Word cloud constructed from the most important words from Cluster-3

Countries in this cluster contain the United States, Canada, Mongolia, Saudi Arabia, Greenland, etc. The theme of freedom is often present in national anthems, as it reflects a country's history of struggles for independence and the values of liberty and self-determination. Many national anthems contain lyrics that celebrate the freedoms and rights enjoyed by the citizens of the country, and express the determination to defend those freedoms against any threats.

For example, the lyrics of the US "The Star-Spangled Banner" describe the flag of the United States and the defense of Fort McHenry during the Battle of Baltimore. The theme of freedom and patriotism are prominent in "The Star-Spangled Banner," as the song celebrates the values of liberty, freedom, and self-determination, and expresses pride in the United States and its flag. Similarly, Canada's national anthem "O Canada" also expresses a similar theme of the country's pride in its history and people, and its commitment to peace, justice, and freedom.

Cluster- 4 - (Fight and protect):



country	anthem	label
France	Arise, children of the Fatherland, The day of ...	type-4
Poland	Poland has not yet died, So long as we still l...	type-4
Cuba	To combat, run, Bayamesans! For the homeland l...	type-4
Bahamas	Lift up your head to the rising sun, March on...	type-4
Barbados	In plenty and in time of need When this fair l...	type-4
Samoa	Samoa, arise and raise your flag, your crown! ...	type-4
China	Arise, ye who refuse to be slaves; With our ve...	type-4
Macau	Arise, ye who refuse to be slaves; With our ve...	type-4
Vietnam	Soldiers of Vietnam, marching onward United in...	type-4
Kenya	O God of all creation Bless this our land and ...	type-4
Niger	Throughout great powerful Niger Which makes na...	type-4
Nigeria	Arise, O Compatriots Arise, O Compatriots, Nig...	type-4
Togo	Hail to thee, land of our forefathers' Thou wh...	type-4

Many national anthems contain themes of fighting, war, and protection. These anthms maybe been were written during times of conflict or struggle, when the country was fighting for its independence or to defend itself against foreign threats. The counties in this cluster include France, Poland, Cuba, Niger, Vietnam etc.

Let’s now compare the national anthem of China and France which are both part of this cluster:

Chinese National Anthem

With our flesh and blood, let us build our
Great Wall!
The Chinese people face their greatest peril.
From each one the urgent call for action
comes forth.
Arise! Arise! Arise!
Us millions with but one heart,
Braving the enemy's fire, march on!
Braving the enemy's fire, march on!
March on! March on, on!
…

French National anthem:

Arise, children of the Fatherland,
The day of glory has arrived!
Against us, tyranny's
Bloody standard is raised, (repeated)
Do you hear, in the countryside,
The roar of those ferocious soldiers?
To cut the throats of your sons, your women
…
To arms, citizens,
Form your battalions,
March, march!
Let an impure blood
Water our furrows!
…

Both of these national anthems share a common theme of war to protect their homeland against foreign enemies.

7.2 Limitation:

The result of the project has given some really insightful explanation of how some of the national anthems share a common theme and can be clustered together. In many cases, the algorithm failed to classify a national anthem to the correct category despite having direct clues. This can be because a national anthem may contain multiple themes and messages that are difficult to capture using machine learning algorithms and NLP techniques. For example, an anthem may contain lyrics that express pride in the country and its people, while also expressing a commitment to peace and justice. These themes may be difficult to accurately identify and classify using machine learning algorithms alone.

Another limitation could be that national anthems often include symbolic language and imagery that may be open to interpretation. This can make it challenging for machine learning algorithms to accurately understand the meaning and significance of the lyrics, and to accurately classify and cluster the anthems based on their themes and messages.

Overall, while machine learning and NLP techniques can be useful for analyzing and clustering national anthems based on their lyrics, it is important to consider these limitations and to approach the task with careful consideration with further improvement in a better word encoding system.

8. Conclusion:

In this research paper, we applied natural language processing techniques and k-means clustering to the lyrics of national anthems from various countries. Our goal was to cluster the anthems into different categories based on their content and themes. We first pre-processed the lyrics, and then applied k-means clustering using a variety of different parameters. Our results showed that it was possible to effectively cluster the anthems into distinct categories based on their themes, such as those focused on unity and patriotism, and celebrate the natural beauty, or tell the story of struggles and war. Overall, our findings demonstrate the potential for using NLP and k-means clustering to analyze and categorize the content of large collections of text data. This type of analysis can provide valuable insights into the themes and messages present in different types of texts.

9. Future Study:

In this study, we analyzed the lyrical characteristics of national anthems from 190 countries and used machine learning techniques to cluster the anthems based on these similarities in the lyrics. There are several directions in which this work could be extended in the future. One potential area of future work is to expand the scope of the analysis to include musical/ or audio analysis of the anthems. Clustering anthems based on audio features like: tempo, pitch, instruments, dancibility, loudness, energy, instrumentalness, acousticness etc could also provide insight into the musical characteristics of different countries. For example, it can be found that certain musical styles or instruments are more common in certain regions or that there are distinct musical "families" that are shared by groups of countries.

11. Code

The code for the project can be found in the following repository:

https://github.com/deepankarck2/National-Anthem-Analysis--Machine-Learning

10. Acknowledgement:

This project has been inspired by the youtube channel ***'India In Pixels by Ashris'*** specifically from the video 'AI classifies anthems into 5 groups'. The video in the original Hindi Language can be found here: https://youtu.be/a-AqvPtjjts,

https://medium.com/@lucasdesa/text-clustering-with-k-means-a039d84a941b

11. Bibliography

Term Frequency — Inverse Document Frequency statistics -

https://jmotif.github.io/sax-vsm_site/morea/algorithm/TFIDF.html

12. Appendix:

Appendix A.

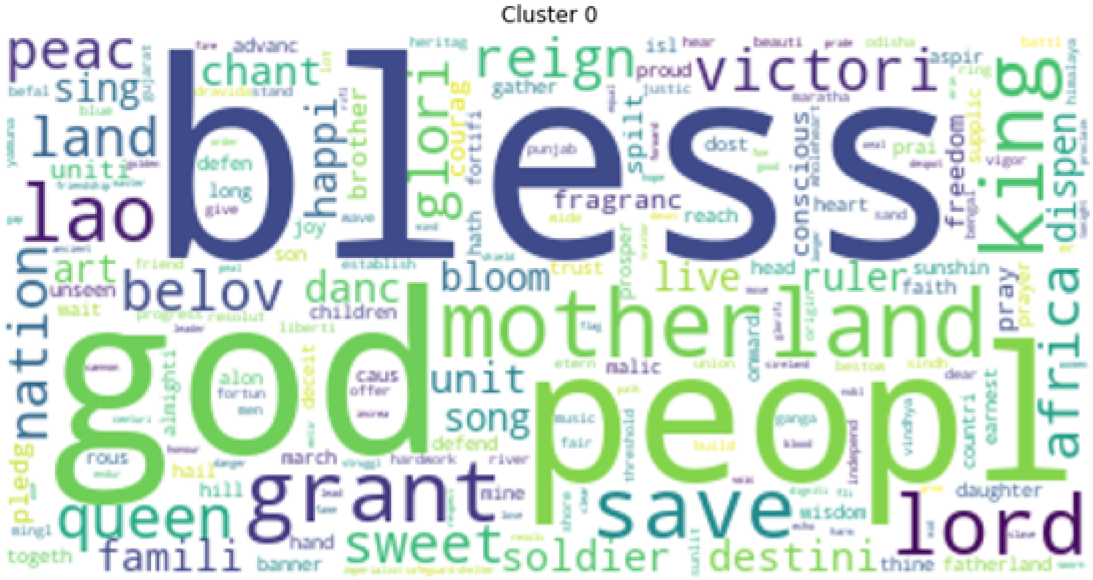


Figure: Word Clould consisting of the most important words in Cluster-0

Appendix B.

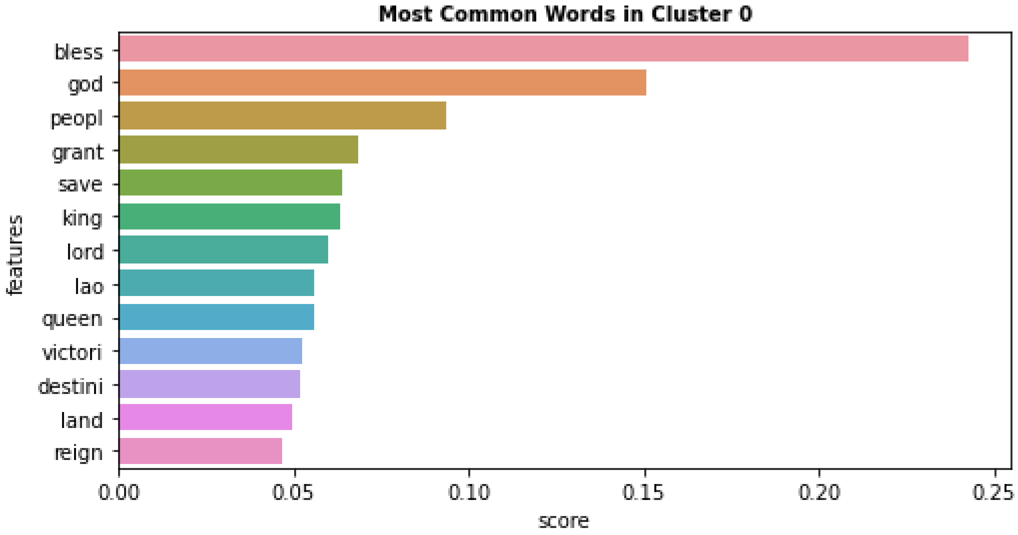


Figure: Word Frequency and importance score in cluster 0

Appendix C.

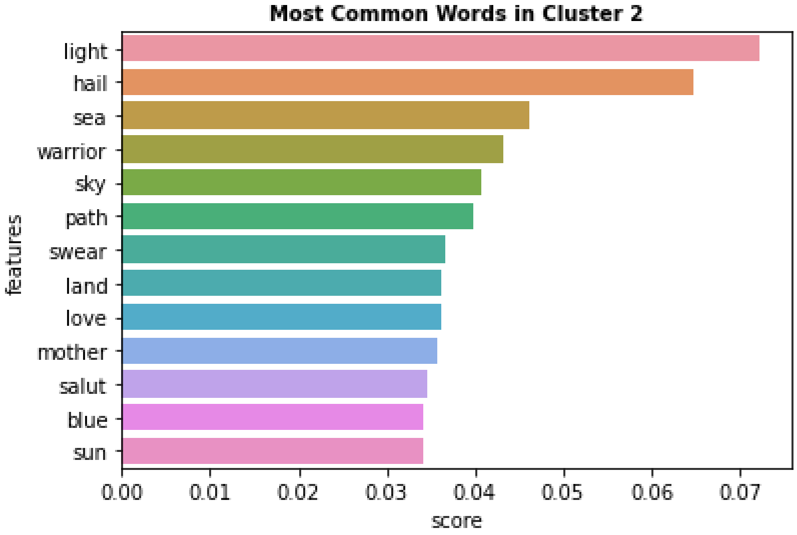


Figure: Word Frequency and importance score in cluster 2

Appendix D.

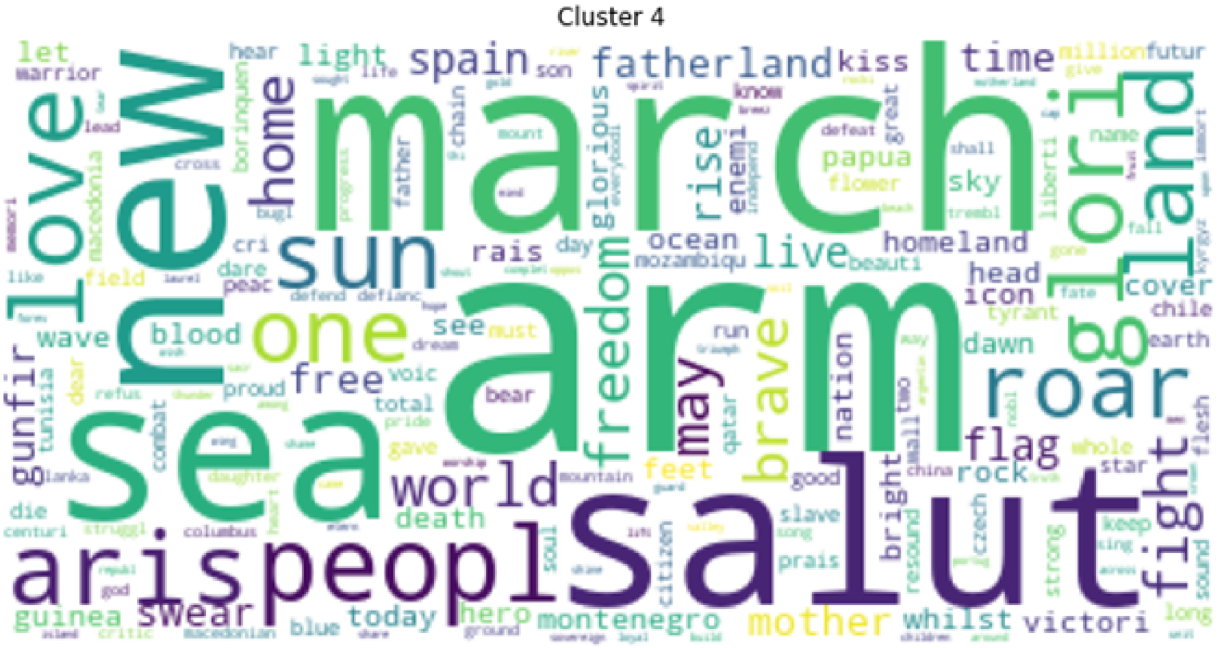


Figure: Word Clould of cluster-4(war and protect)