CS 410 Tech Review
Christopher Dimitri Sastropranoto 678097021

K means clustering is an unsupervised machine learning algorithm that has many applications in industry today. One such application is document classification, where a large number of documents are fed to the algorithm and categorized. This review will provide a brief background of how the algorithm works and how it can be applied to a real world problem. The example used in this paper will be to apply clustering to categorize national anthems into separate categories.

K Means Clustering

The main goal of k means clustering is to take a set of data points and group them into separate clusters. Points grouped in the same clusters are then said to be in the same class. Two important metrics for the algorithm are the distance and error functions. The distance function measures how far two points are whilst the error function measures how good the clusters are.

Algorithm Steps:

1. Take k random points and designate them as the original cluster centers.
2. Take the distance of each point to each of the k clusters.
3. Assign each point to the cluster ID of the closest cluster center.
4. Compute the new cluster averages and assign that to be the new center.
5. Measure the error function.
6. Repeat steps 2 - 5 until the improvement to the error function is below a preselected threshold.

National Anthems Example

An example application of k means is to feed it a set of national anthems and classify them into different categories depending on their lyrics. Categories range from militaristic anthems to more patriotic anthems. A link to the dataset is provided in the appendix below.

Preprocessing

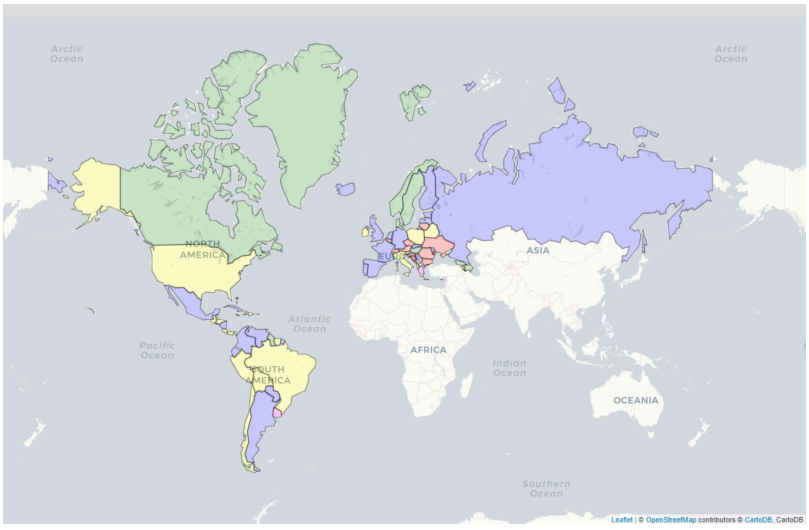
Unfortunately the national anthems dataset cannot be directly used as input to the K means clustering algorithm. This is because the algorithm requires the input data to be in a vector format. The preprocessing steps are outlined below.

1. **Tokenization:** The lyrics are split up into individual words.
2. **Stop Word Removal:** Stop words are common words that provide little information when doing text analysis and can be removed. These are words such as *a, the, not, etc.*
3. **Noise Removal:** Remove non English words. This includes removing numbers and other ASCII characters such as ? And !.
4. **Stemming:** Reduce individual words to it's root form.
5. **Feature Extraction:** Map documents into a word vector. Each entry in the vector is the TF-IDF weighting of the word.

Running the Algorithm

The results of the preprocessing step can then be fed into the algorithm to obtain a set of clusters. Each cluster represents anthems that belong to the same category based off it's lyrics. For this analysis, it was found that the optimal value number of clusters was 5.

Results



Red - Anthem has lyrics praising motherland.

Yellow - Words praising liberty.

Green - Religious anthems.

Blue - Warlike lyrics.

Magenta - Generic lyrics and were not well clustered.

Conclusion

In conclusion K means clustering is a powerful algorithm that can be used to find patterns in unstructured text data. Whilst the example provided applied that algorithm on a trivial case, the algorithm can be applied to solve more complex situations such as sentiment analysis.

Acknowledgements

https://medium.com/bexs-test/text-clustering-with-k-means-a039d84a941b