Report

Introduction

The data given to us in this assignment is the normalized frequency of function words used in the Federalist papers. The authors are Hamilton (51), Madison (15), Hamilton and Madison (3) and Jay(5). There are 11 unidentified essays which could be written by Hamilton or Madison. The data provided has columns for authors and has 11 rows where the authors are disputed.

Data Preprocessing and analysis

The data has 5 unique authors 'dispt', 'Hamilton', 'HM', ‘Jay' and ‘Madison’. Since we want to see what cluster the disputed papers (author: ‘dispt’) fall under, we will use 4 clusters.

There are no NA values in the dataset.

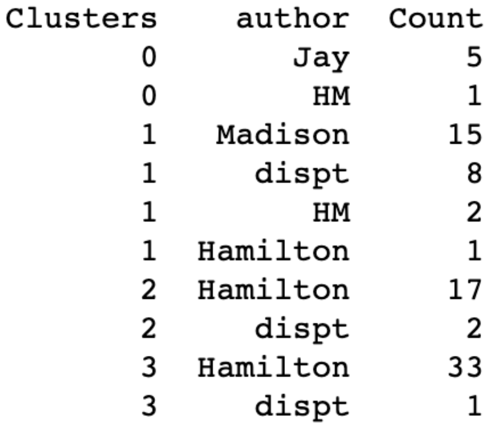
For scaling I created a function which checks the minimum and maximum values of each column and prints the column names that need to be scaled and then proceeds to scale them using the MinMaxScaler from sklearn. The two columns ‘of’ and ‘the’ were not properly scaled.

K-Means Clustering

K-means is an unsupervised machine learning algorithm used for clustering data points into groups or clusters based on their similarity. The algorithm works by iteratively partitioning the data into K clusters, where K is a predetermined number of clusters specified by the user.

I have chosen our k to be 4 based on the number of authors in our dataset. Even though problem statement only asks us to make distinctions between the two authors Madison and Hamilton, taking 4 clusters instead of 2 gave significantly clearer results.

Joining the labels (clusters) with the original table and summarizing the data based on the author and the cluster they fall under. We get the following table



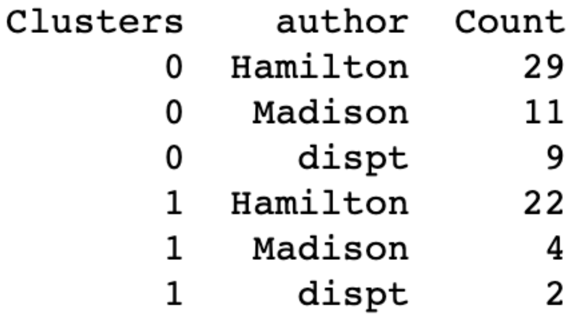
```
Clusters    author  Count
       0       Jay      5
       0        HM      1
       1   Madison     15
       1     dispt      8
       1        HM      2
       1  Hamilton      1
       2  Hamilton     17
       2     dispt      2
       3  Hamilton     33
       3     dispt      1
```

This table reveals that cluster 0 belongs to Jay. The author Jay does not appear in any clusters. One paper which was written by both Hamilton and Madison has been misclustered into this category.

We can see that Cluster 1 is dominated by Madison and most of the disputed papers fall into this cluster. One paper from Hamilton has been grouped with this cluster. HM is papers written by both Hamilton and Madison so it makes sense that the kmeans clustering algorithm finds it difficult to distinguish between the two authors.

In this Case cluster 2 and 3 both have Hamilton as the dominating author with a few disputed papers also falling into the clusters.

Doing this analysis with 2 clusters instead of 4 after removing the papers authored by Jay and HM I got worse results. The table below shows 29 papers by Hamilton in cluster 0 and 22 papers in cluster 1. Thus we can not make a clear distinction as to which cluster belongs to which author.



```
Clusters    author  Count
       0  Hamilton     29
       0   Madison     11
       0     dispt      9
       1  Hamilton     22
       1   Madison      4
       1     dispt      2
```

Conclusion

In conclusion we can say that disputed papers were most likely written by Madison. The ideal number of clusters to analyze this data is 4.