Name: Prachi Sadarangani

Date: 09/29/2023

**Who wrote the disputed essays, Hamilton or Madison?**

The aim of this report is to find the authorship of the disputed essays from the Federalist Papers dataset. The dataset contains 51 essays attributed to Hamilton, 15 to Madison, three to both Hamilton and Madison, 5 to John Jay, and 11 essays that are disputed. The goal is to use k-means clustering and Hierarchical Agglomerative Clustering (HAC) to determine who wrote the essays, Hamilton or Madison.

There are multiple features in this dataset so before performing the clustering, we will reduce the dimension. The dataset has been standardized and divided into X and Y where X is the dataset with the features and without the target variable whereas Y is the target variable, which is the author in this case. Then PCA has been performed on X, the n components are not specified so that we can retain all the features. The below graph shows the scatter plot for PCA with the authors as the target and labels.

A chart with colored dots

Description automatically generated

Now, this PCA data will be used for the clustering. Let’s start with K-means clustering. The cluster size chosen for this is 4 because there are 4 authors. The idea is to divide the data into 4 clusters and compare the clustered data labels with the original authors. The below graph shows the clustered data with labels.

A diagram of a number of dots

Description automatically generated with medium confidence

In the original data frame, a new column is added called the ‘kmeans labels’. Now, the data frame is filtered to get only the disputed papers. From the below picture, it is visible that 9/10 disputed papers have the ‘kmeans labels’ as 3 which means they belong to cluster 3.

A table with numbers and lines

Description automatically generated

Now, let’s filter the original data frame where ‘kmeans labels’ is equal to 3 and the author is not ‘dispt’ to see which other authors are in cluster 3. The below picture shows the count of the authors in cluster 3. Madison has 15 papers in cluster 3 which is the maximum. Hence, this evidence points to the conclusion that Madison is the author.

A close-up of a number

Description automatically generated

Now, let’s proceed with HAC. HAC is performed on the PCA dataset using the linkage function. To assign labels to the clusters, the fcluster function is used and the number of clusters is set to 4 again because there are 4 authors. The dendrogram is shown below.

A diagram of a city

Description automatically generated

Now, the same method as above is used, a new column is added in the original data frame called the ‘cluster labels’. Now, the data frame is filtered to get only the disputed papers. The below picture shows that 9 out of 10 disputed papers are in cluster 3.

A screenshot of a table

Description automatically generated

Now, the data is filter by choosing the cluster label as 3 and where the author is not ‘dispt’. The below picture shows the output.

A table with numbers and symbols

Description automatically generated

It is visible that in cluster 3, there is only one author who is Madison. The disputed papers belong to this cluster.

Hence, using the K-means and HAC clustering, we can conclude that Madison is the author of the disputed papers.