Using Clustering to determine the real author of the disputed papers

Hamilton or Madison?

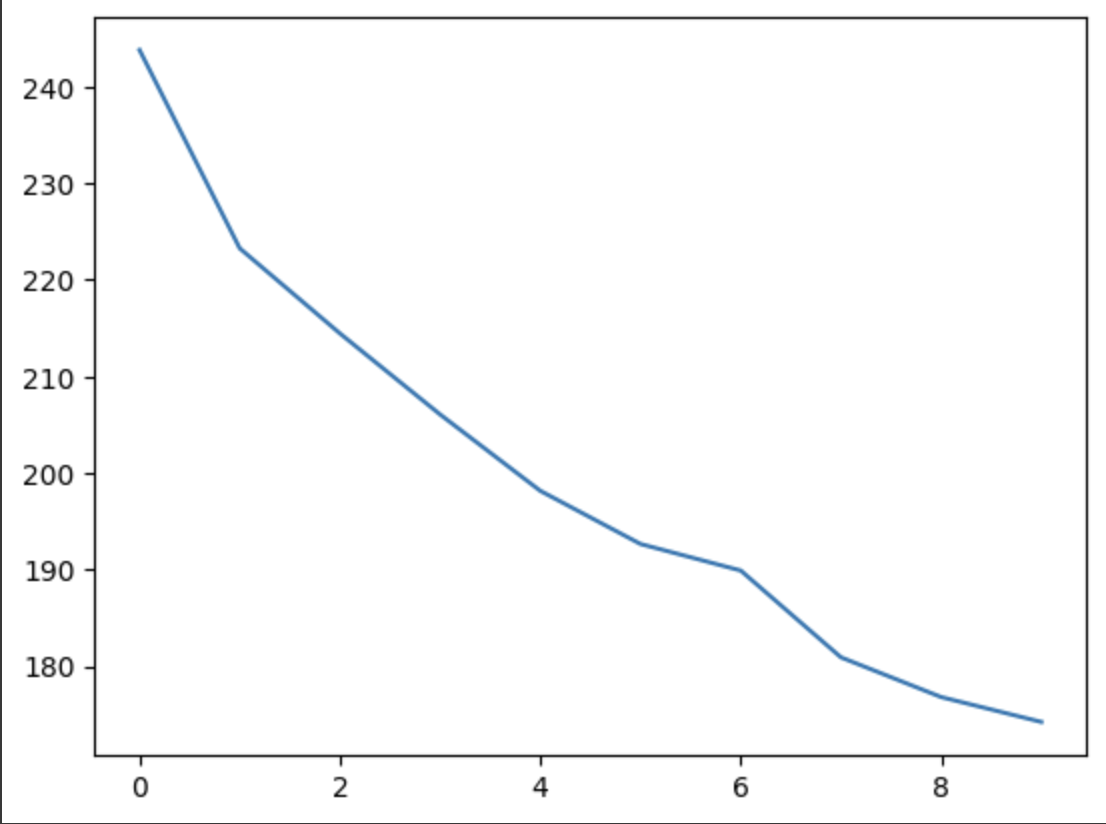
Vipul Rajiv Sarode

**Introduction**

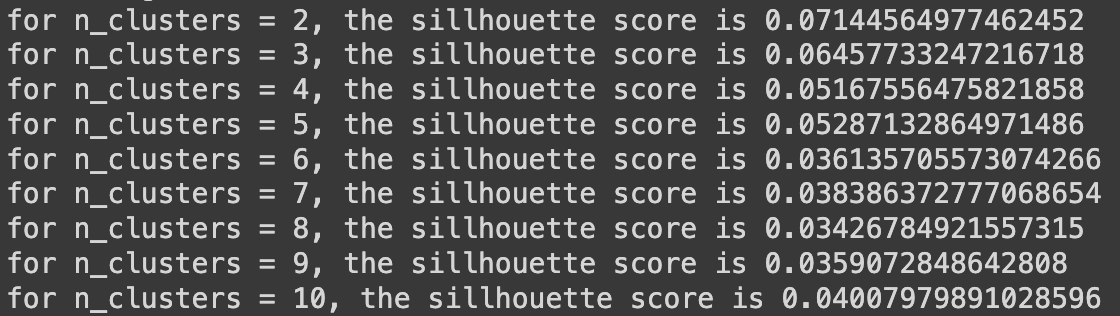
We have a dataset that contains text embedding of words used by famous authors in their papers. The data consists of 3 authors named Jay, Hamilton, Madison. Other two categories include the papers written by Hamilton and Madison collectively and the papers who are in question, whether they had been written by Hamilton or Madison. We will be using KMeans Clustering to determine if the papers are written by Hamilton or Madison.

**Analysis**

As the dataset we have has a lot of words that has zero usage, and many other words have ample amount of usage, hence we used MinMaxScaler to normalize the data. The data has no null values. The data has high dimensionality with 70 features, so we reduced the number of features to 42, which explains 95% of the variance in the data, using Principal Component analysis. Next, we use the Elbow plot method and silhouette scores to determine the number of clusters to be considered while performing the clustering.



**Fig 1 – Elbow Plot**



**Fig 2 – Silhouette Scores per number of clusters.**

By observing the elbow plot, silhouette scores and the problem at hand, we set the number of clusters to be predicted to 5. Hence, we trained the Kmeans unsupervised algorithm by considering 5 clusters.

.

**Results**

A screenshot of a computer

Description automatically generated

**Fig 3 – Number of occurrences of an author per cluster**

As we can see from Fig 3, the cluster 4 contains the majority of Madison – written and disputed papers, which hints us that these papers are very closely related. Hence, we conclude from our analysis that Madison is the real author of the disputed federal papers.