HW 04 – CLUSTERING

**DATA**

Dataset is a series of 85 federalist papers. Data is tokenized and provided in CSV format. Tokens are function words/feature set with feature value as percentage of word occurrence in the essay. Data is loaded as data frame having 85 rows and 72 columns. Each row represents an essay, and each column represents function words.

Graphical user interface, application, Word

Description automatically generated

Viewing first few rows of the data frame:

A picture containing table

Description automatically generated

Frequency distribution of Essays: Most of the essays are written by Hamilton.

Chart, pie chart

Description automatically generatedGraphical user interface, application

Description automatically generated with medium confidence

There are no Null values or NAs present in the data.

Chart, bar chart

Description automatically generated

A picture containing shape

Description automatically generated

**K – Means Clustering**

From the elbow method, optimal number of clusters found to be 5.

Chart, line chart

Description automatically generated

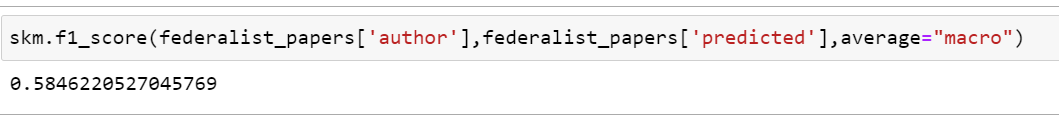
Ran the K-Means clustering algorithm for creating 5 clusters on the federal dataset. Silhouette score of the clusters formed by the algorithm is 0.12.

Chart

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

F1 score of the classification obtained is 0.58



**Hierarchical Clustering**

From the dendogram, optimal number of clusters found to be 3 but according to given dataset, 5 clusters are possible [Hamilton, Jay, Hamilton and Madison, Madison, and Disputed].

Chart

Description automatically generated

Ran the Hierarchical clustering algorithm for creating 5 clusters on the federal dataset. F1 score of the classification obtained is 0.69

Text

Description automatically generated with low confidence

**CONCLUSION**

Chart, pie chart

Description automatically generated

1. From the Authors pie chart, we can notice that **data is imbalanced**. Data set contains **60%** of the federal papers written by **Hamilton** and only **19%** of the federal papers are written by **Madison**. Since the dataset has more of Hamilton essays, results of the clustering algorithm can be biased.
2. There is not enough data to **perform random sampling to balance the dataset.**  With only 18% of Madison essays available, if we sample the dataset, we will be further reducing the training dataset

**K - Means Clustering Analysis**

1. From the **elbow** method, **optimal number of clusters found to be 5**

2. **Silhouette score** of the K-means model is **0.12**: Silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. 0.12 seems to be a good score; clusters are located far apart

3. **Confusion Matrix:** As per the confusion matrix,

a. Hamilton’s essays are being grouped in cluster #2 and #3

b. Jay’s essays are being grouped in cluster #0

c. Madison’s essays are being grouped in cluster #1, #4

A picture containing table

Description automatically generated

According to the prediction obtained by K-Means clustering model, most of the disputed essays are written by Madison however few of the disputed essays are written by Hamilton, and few are written by both.

Heat Map of Confusion Matrix from K-Means Clustering

A picture containing icon

Description automatically generated

**Hierarchical Clustering**

**Chart

Description automatically generated**

From the dendogram, we were able to find optimal number of clusters as 3 but, we already know the number of categories of the essays – Hamilton, Jay, Madison, Hamilton & Madison, disputed. So, number of **clusters considered in hierarchical clustering is 5**

**A picture containing table

Description automatically generated**

1. **Confusion Matrix:** As per the confusion matrix,

a. Hamilton’s essays are being grouped in cluster #0 and #2

b. Jay’s essays are being grouped in cluster #3

c. Madison’s essays are being grouped in cluster #1, #2

According to the prediction obtained by Hierarchical clustering model, half of the disputed essays are written by Madison and other half by Hamilton.

Heat Map of Confusion Matrix from Hierarchical Clustering

**A picture containing logo

Description automatically generated**

Both K-Means clustering, and Hierarchical clustering indicate that few of the disputed essays are written by Hamilton, and few are written by Madison. However, it is not co-authored. In both the clustering models, essays with joint authorship are grouped as a separate cluster indicating they have very different style of writing when they co-author which is quite different when they write the essays individually. Most of the disputed essays are clustered either in Hamilton or Madison but not in the cluster of Hamilton and Madison. Clearly, they did not co-author the disputed essays.