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

Chart, bar chart

Description automatically generatedThere are no Null values or NAs present in the data. Hence, data cleaning is not needed.

A picture containing shape

Description automatically generated

**EXPLORATORY ANALYSIS**

Frequency distribution of Essays based on authors.

Chart, pie chart

Description automatically generatedGraphical user interface, application

Description automatically generated with medium confidence

Most of the essays are written by Hamilton.

**Finding pattern in the verbiage of each author:**

**Madison Essays**

**Background pattern

Description automatically generated**

Hamilton Essays

A picture containing text

Description automatically generated

Disputed Essays:

Background pattern

Description automatically generated with medium confidence

It was very difficult to find any pattern in the essays. Most of the words given in the CSV file seems to be stop words.

Maybe we need to analyze the raw text files of all the papers to find out any specific word pattern between the authors.

**K – Means Clustering**

Chart, line chart

Description automatically generatedFrom the elbow method, we can conclude that optimal number of clusters can be between 3 and 5. But, with the domain knowledge, we can conclude 4 clusters – Hamilton, Madison, HM, and Jay.

Chart

Description automatically generatedK-Means clustering algorithm with 4 clusters on the federal dataset resulted in Silhouette score of 0.11.

Text

Description automatically generated

Silhouette score varies between -1 and +1. Closer the Silhouette score to +1, better the clustering. So, 0.11 seems to be a good score but, still there is a room for improvement.

**Hierarchical Clustering**

From the dendogram, optimal number of clusters found to be 4. The domain knowledge that we have also suggests 4 clusters - Hamilton, Madison, HM, and Jay Chart, histogram

Description automatically generated

**CONCLUSION**

**K - Means Clustering Analysis**

1. From the **elbow** method, **optimal number of clusters can be between 3 and 5.** According to the domain knowledge that we have, 4 clusters can be created – Hamilton, Jay, Madison, and HM

2. **Silhouette score** of the K-means model is **0.11**: 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. 0.11 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 #1 and #3

b. Jay’s essays are being grouped in cluster #2

c. Majority of Madison’s essays are being grouped in cluster #0

Table

Description automatically generated with medium confidence

According to the K-Means clustering, **most of the disputed essays are written by Madison** however there are couple of disputed essays are being clustered under Hamilton.

Heat Map of Confusion Matrix from K-Means Clustering

A picture containing background pattern

Description automatically generated

**Hierarchical Clustering**

1. From the dendogram, we were able to find optimal number of clusters as 4. The domain knowledge that we have also suggests the same. So, number of **clusters considered in hierarchical clustering is 4**
2. **Confusion Matrix:** According to 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 #0, #1

**Table

Description automatically generated**

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 icon

Description automatically generated**

In both K-Means and Hierarchical clustering there is no clear segregation of disputed essays to either Hamilton or Madison clusters. However, K-Means clustering to some extent indicate most of the essays are written by Madison.