Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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.

```
# loading the data set
federalist_papers = pd.read_csv('HW4-data-fedPapers85.csv') # j
federalist_papers.shape # 85 rows and 72 columns in the data se
```

```
(85, 72)
```

Viewing first few rows of the data frame:

```
federalist_papers.head() # viewing the first 5 rows in the data set
```

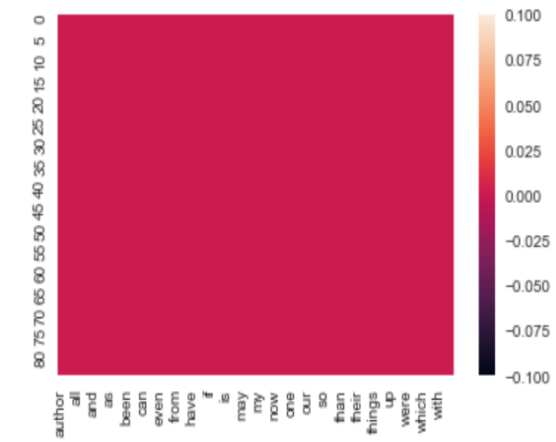
	author	filename	a	all	also	an	and	any	are	as	...	was	were	what	when	which	who	will	with	would	your
0	dispt	dispt_fed_49.txt	0.280	0.052	0.009	0.096	0.358	0.026	0.131	0.122	...	0.009	0.017	0.000	0.009	0.175	0.044	0.009	0.087	0.192	0.0
1	dispt	dispt_fed_50.txt	0.177	0.063	0.013	0.038	0.393	0.063	0.051	0.139	...	0.051	0.000	0.000	0.000	0.114	0.038	0.089	0.063	0.139	0.0
2	dispt	dispt_fed_51.txt	0.339	0.090	0.008	0.030	0.301	0.008	0.068	0.203	...	0.008	0.015	0.008	0.000	0.105	0.008	0.173	0.045	0.068	0.0
3	dispt	dispt_fed_52.txt	0.270	0.024	0.016	0.024	0.262	0.056	0.064	0.111	...	0.087	0.079	0.008	0.024	0.167	0.000	0.079	0.079	0.064	0.0
4	dispt	dispt_fed_53.txt	0.303	0.054	0.027	0.034	0.404	0.040	0.128	0.148	...	0.027	0.020	0.020	0.007	0.155	0.027	0.168	0.074	0.040	0.0

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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

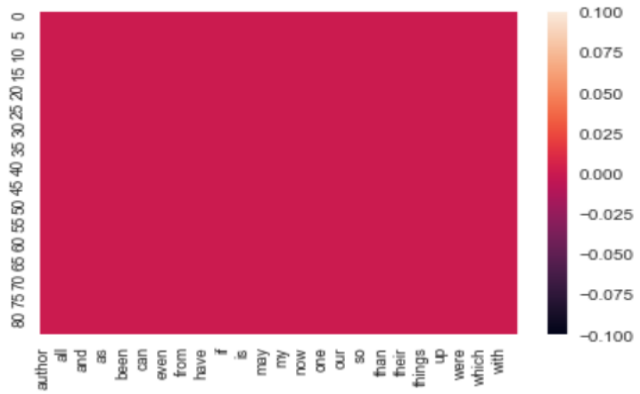
```
sns.heatmap(federalist_papers.isnull()) # no null values present
```

```
<AxesSubplot:>
```




```
sns.heatmap(federalist_papers.isna()) # no NA's are present
```

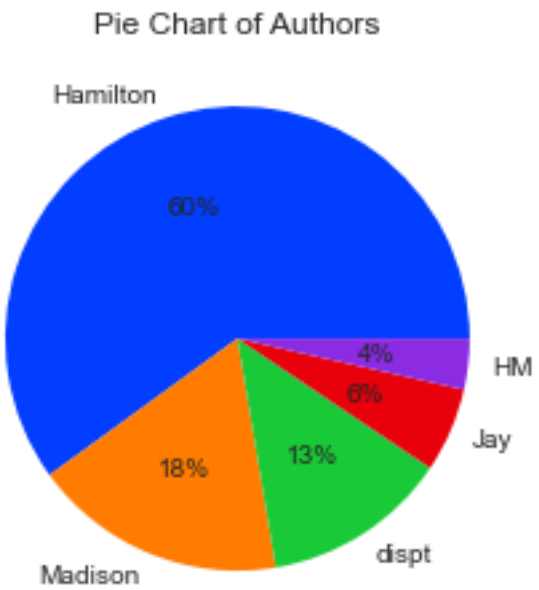
```
<AxesSubplot:>
```



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

EXPLORATORY ANALYSIS

Frequency distribution of Essays based on authors.




```
#Summary of the authors
federalist_papers[['author']].describe()
```

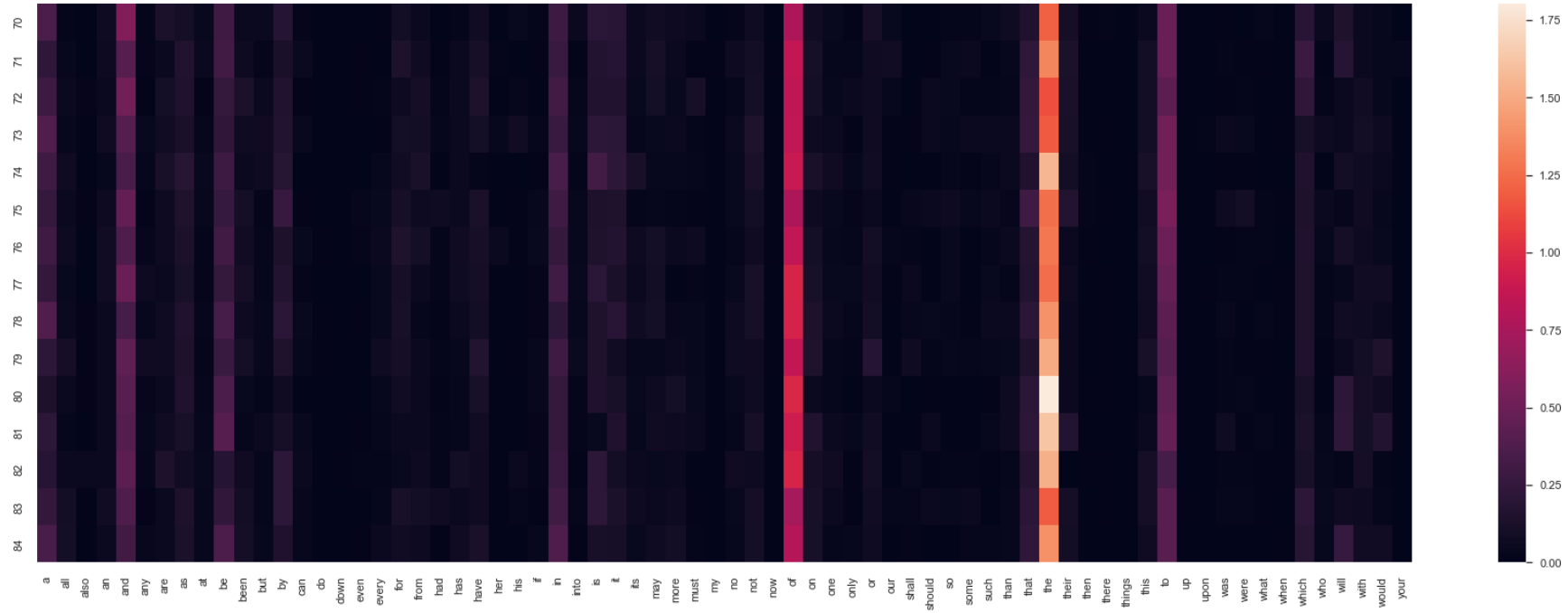
	author
count	85
unique	5
top	Hamilton
freq	51

Most of the essays are written by Hamilton.

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

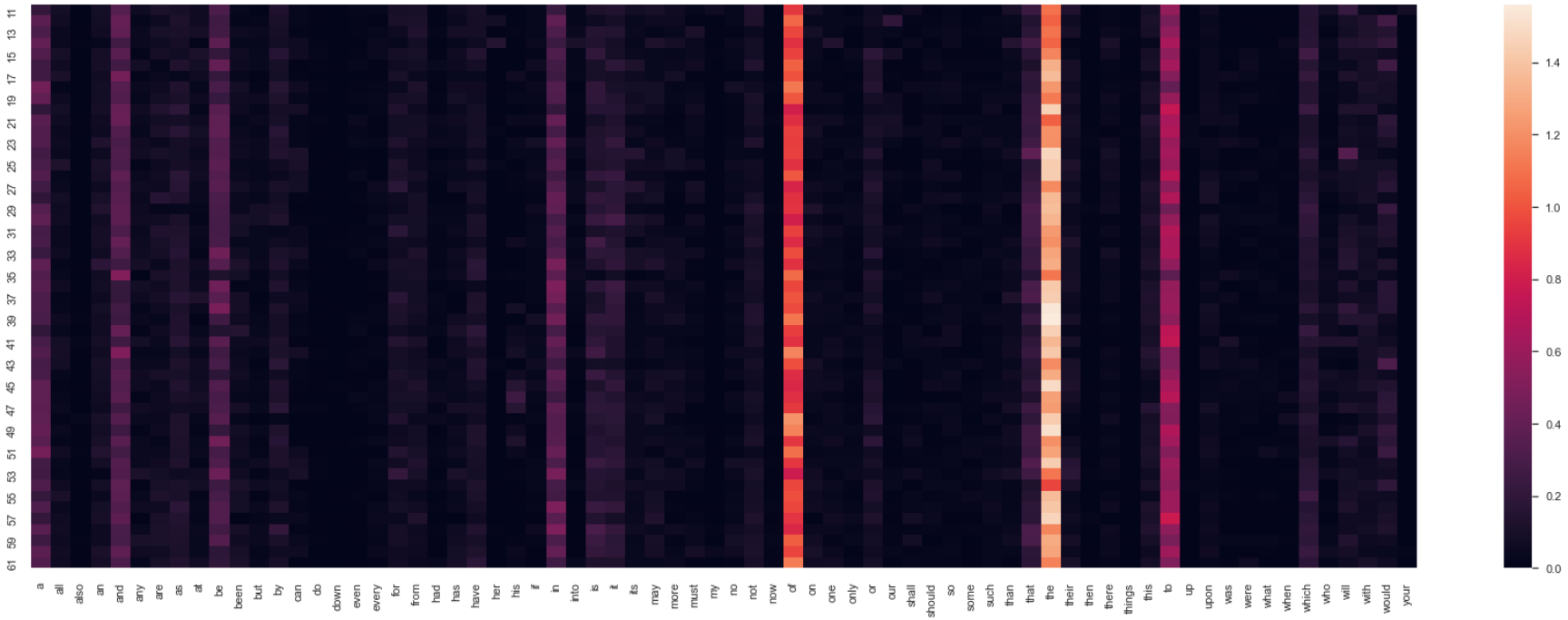
Finding pattern in the verbiage of each author:

Madison Essays



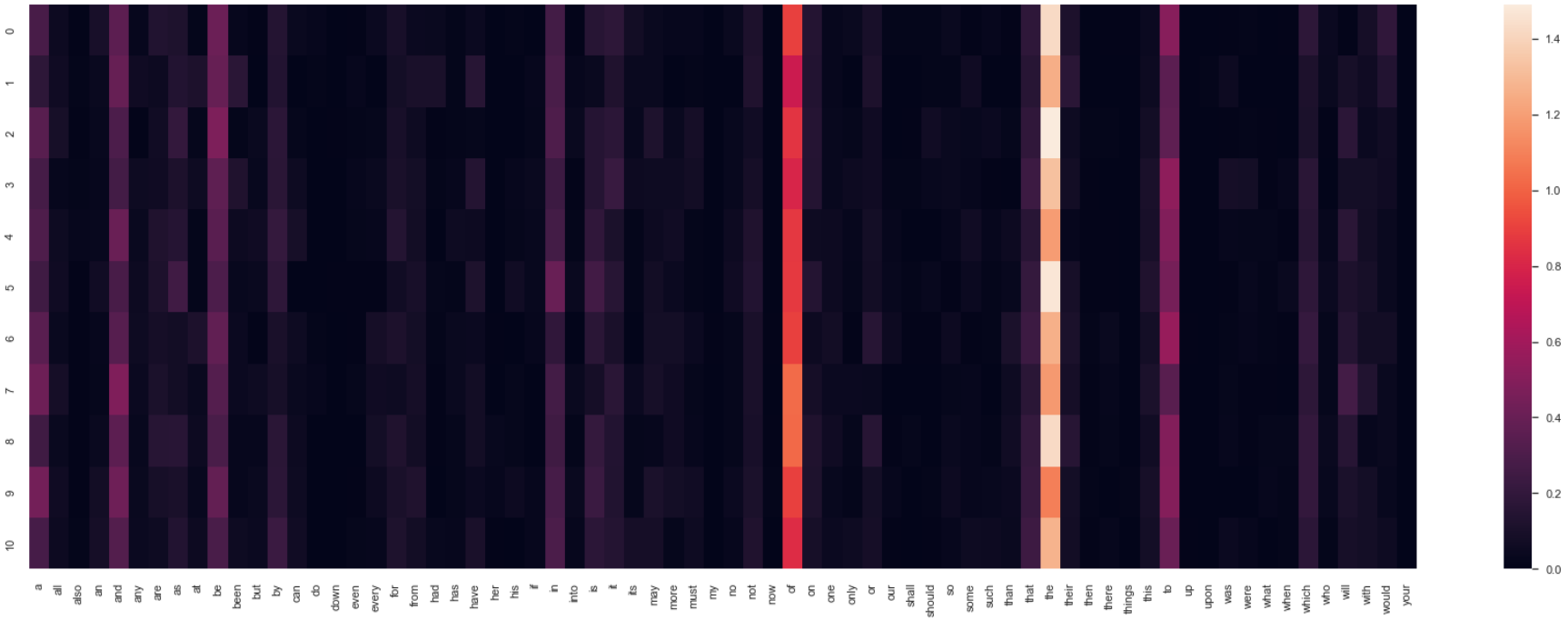
Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

Hamilton Essays



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

Disputed Essays:



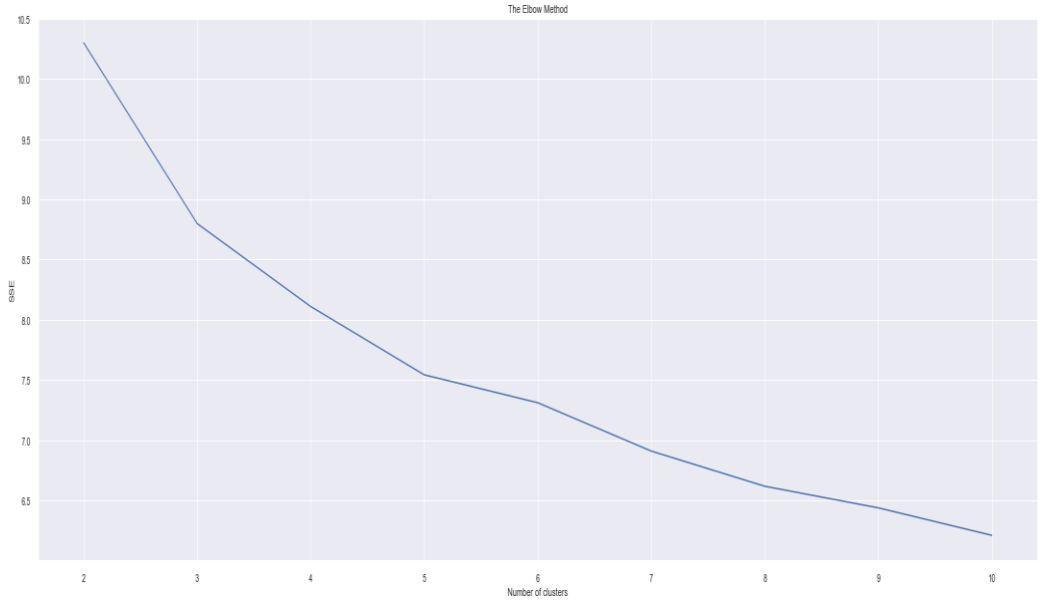
It was very difficult to find any pattern in the essays. Most of the words given in the CSV file seems to be stop words.

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

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

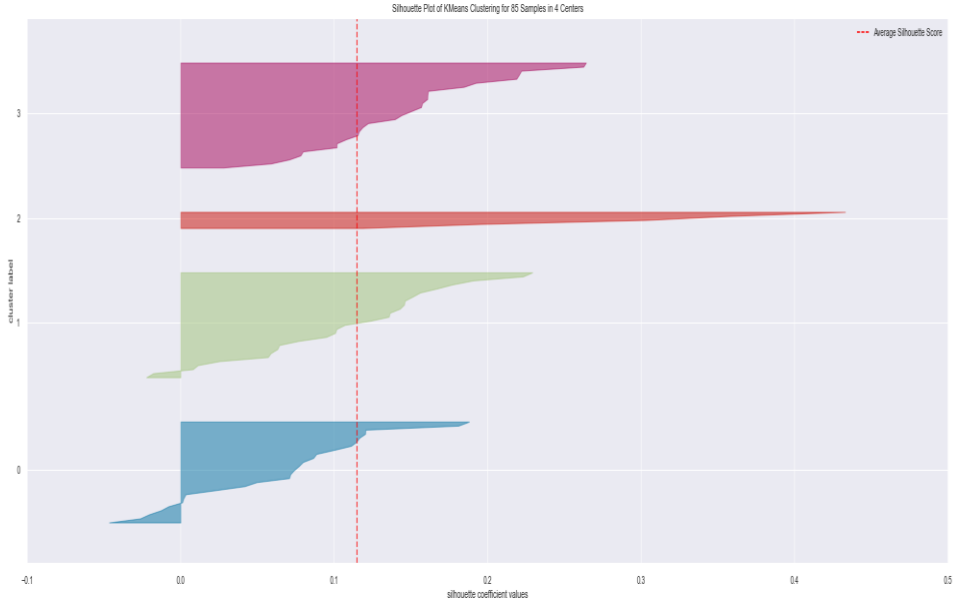
K – Means Clustering

From 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.



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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




```
score = silhouette_score(paper_arr, y_kmeans)
score # 0.11 is still a good score because score is not negative.
```

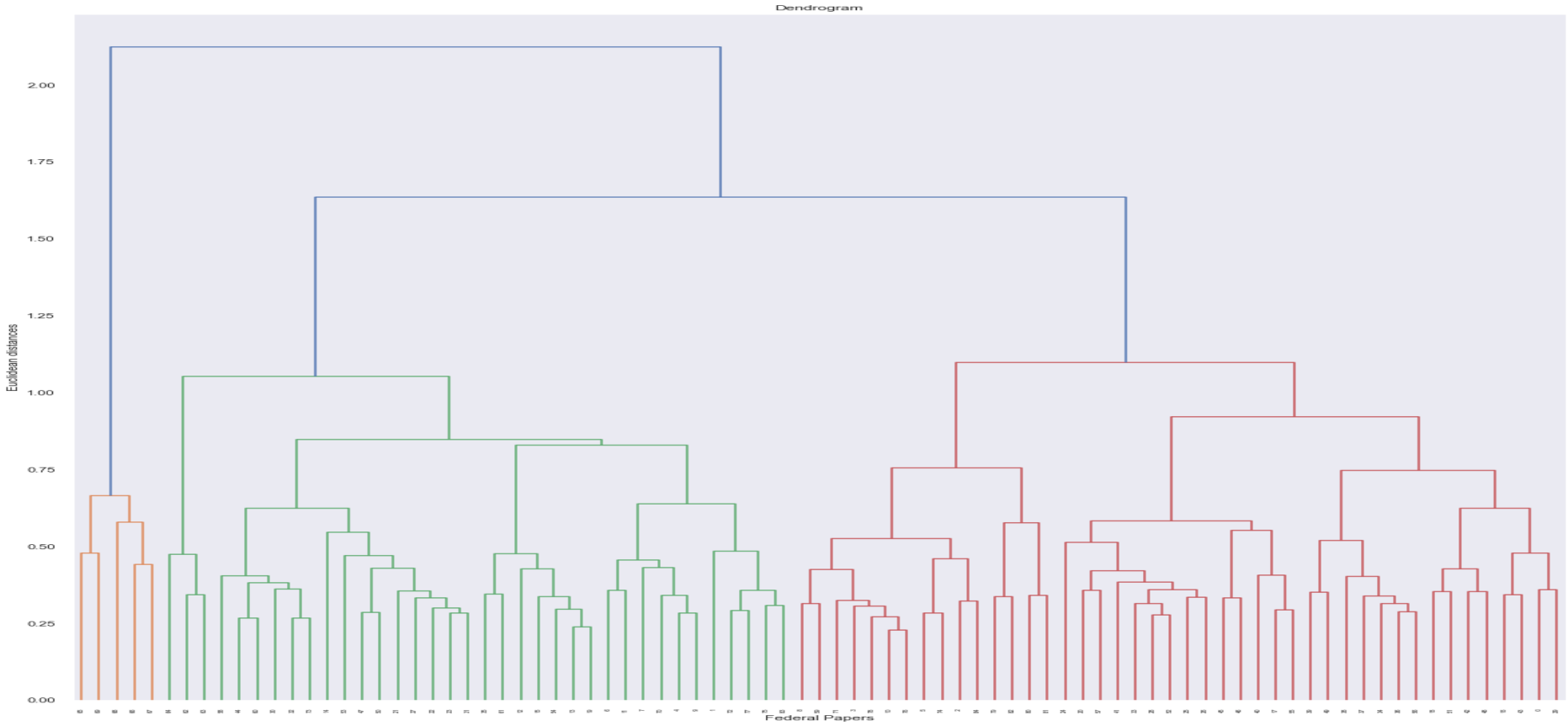
```
0.11494010838022752
```

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.

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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

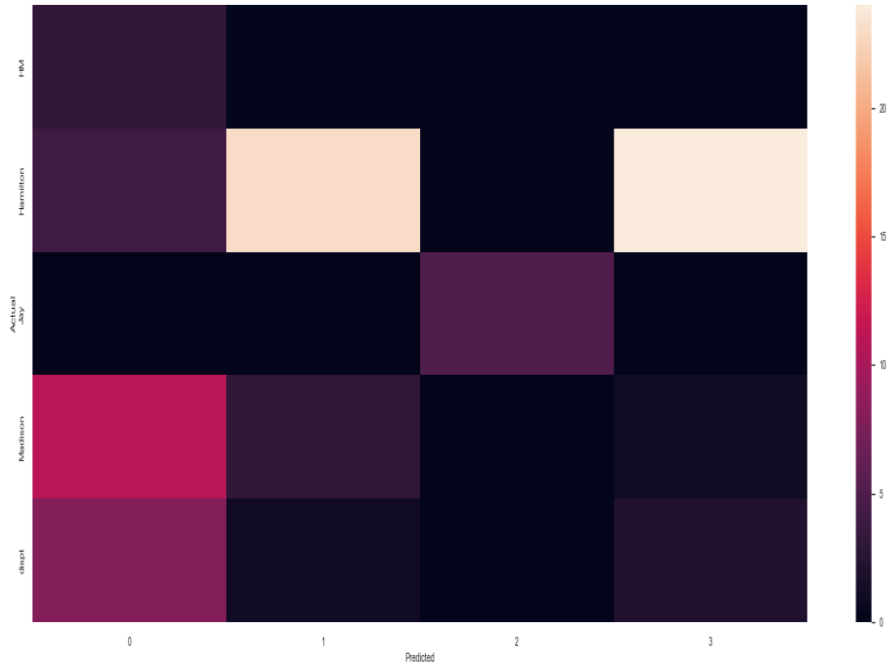
   c. Majority of Madison's essays are being grouped in cluster #0

Predicted / Actual	0	1	2	3
HM	3	0	0	0
Hamilton	4	23	0	24
Jay	0	0	5	0
Madison	11	3	0	1
dispt	8	1	0	2

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.

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

Heat Map of Confusion Matrix from K-Means Clustering



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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**

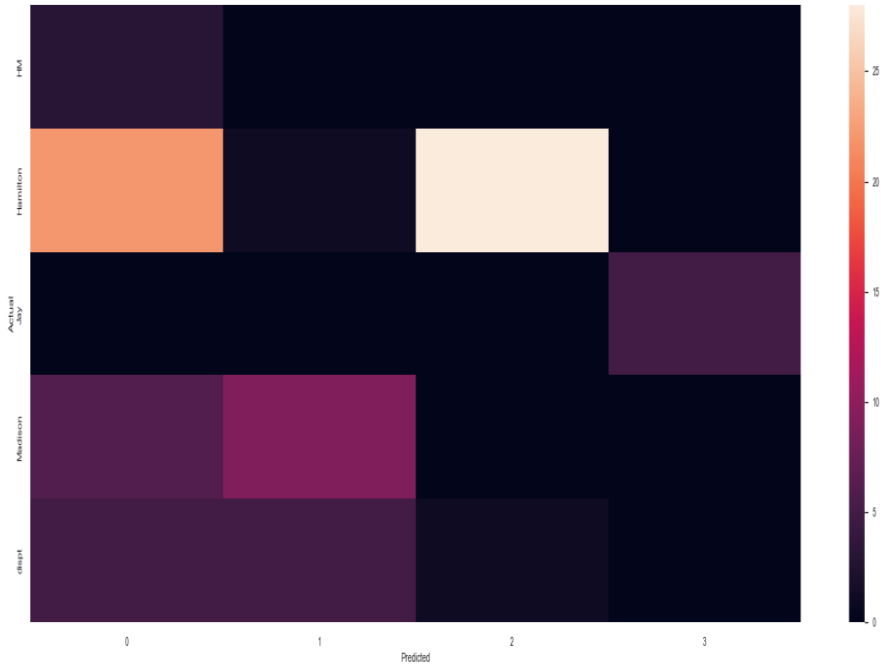
2. **Confusion Matrix:** According to confusion matrix,
   a. Hamilton's essays are being grouped in cluster #0 and #2
   b. Jay's essays are being grouped in cluster #3
   c. Madison's essays are being grouped in cluster #0, #1

Predicted / Actual	0	1	2	3
HM	3	0	0	0
Hamilton	22	1	28	0
Jay	0	0	0	5
Madison	6	9	0	0
dispt	5	5	1	0

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

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

Heat Map of Confusion Matrix from Hierarchical Clustering



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.