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)
```

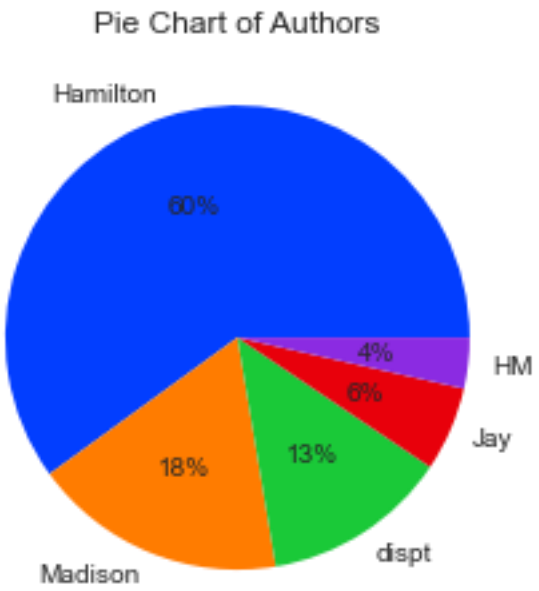
Viewing first few rows of the data frame:

```
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

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

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu




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

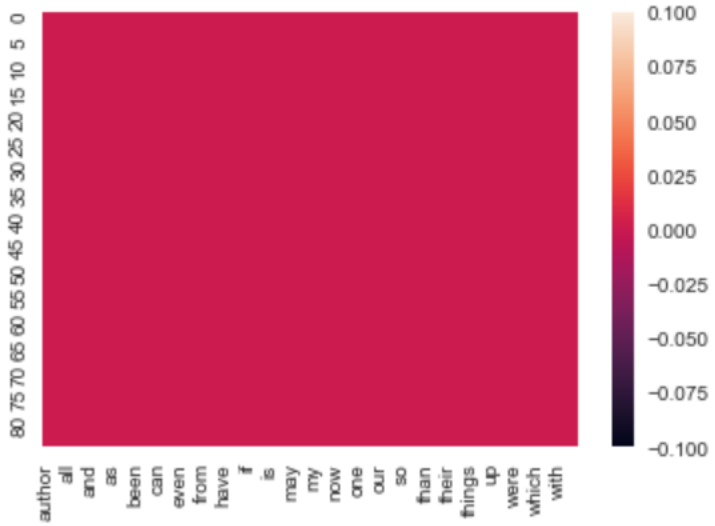
	author
count	85
unique	5
top	Hamilton
freq	51

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

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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

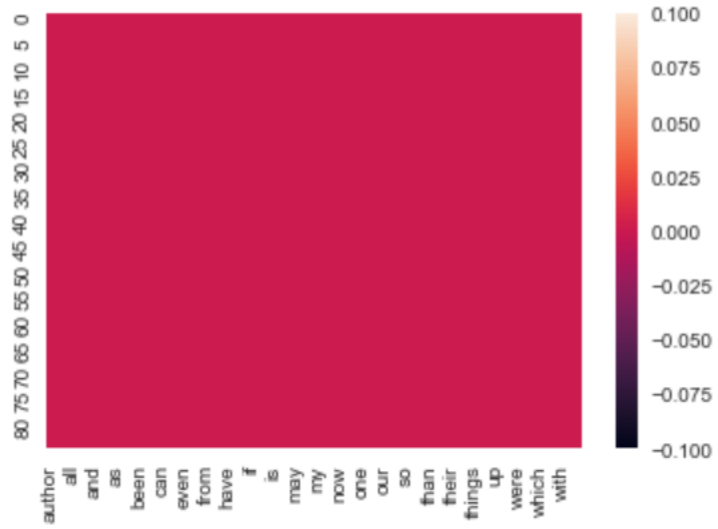
```
<AxesSubplot:>
```



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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

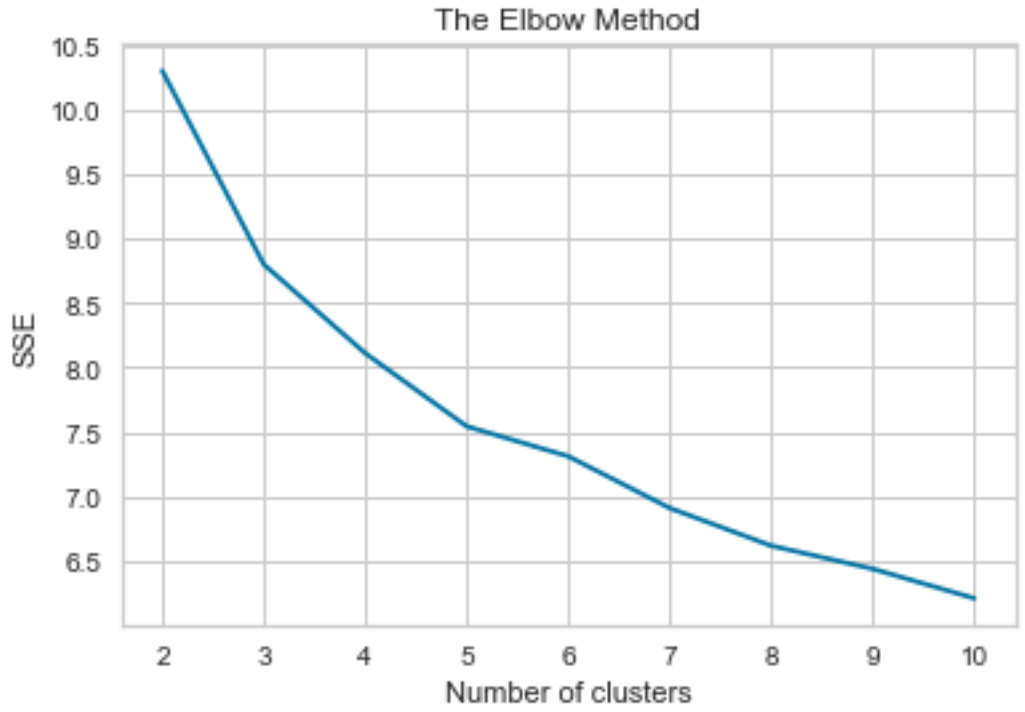
```
<AxesSubplot:>
```



Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

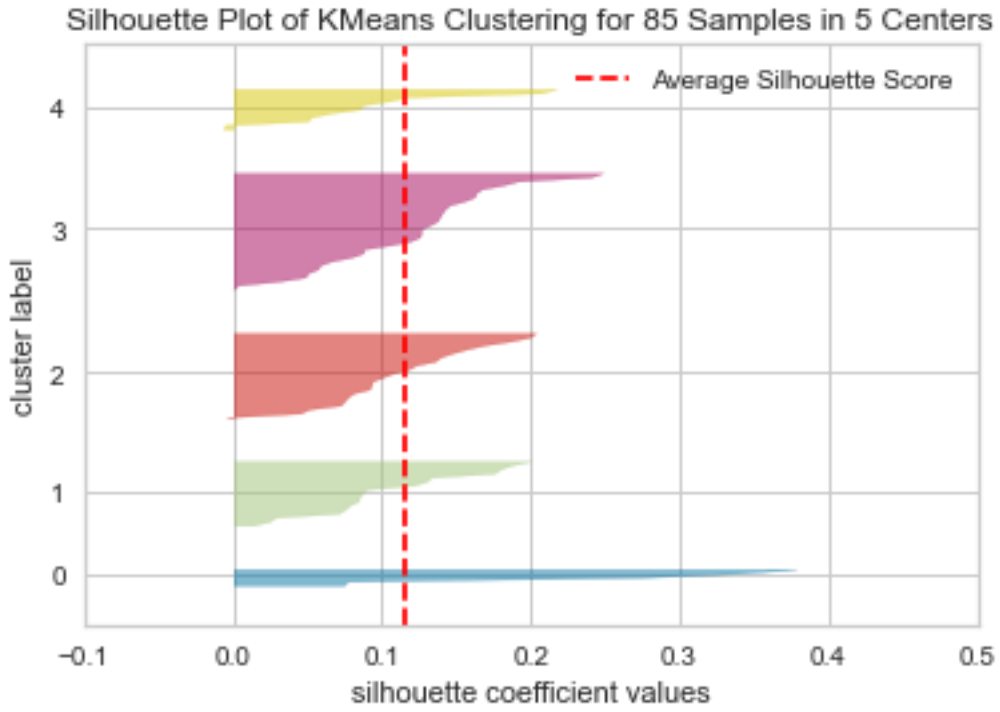
K – Means Clustering

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



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.

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu



VALIDATING THE CLUSTER MODEL WITH SILHOUETTE SCORE

```
: score = silhouette_score(paper_arr, y_kmeans)
  score # 0.12 is still a good score because score is not ne
```

```
: 0.11557311610199418
```

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

F1 score of the classification obtained is 0.58

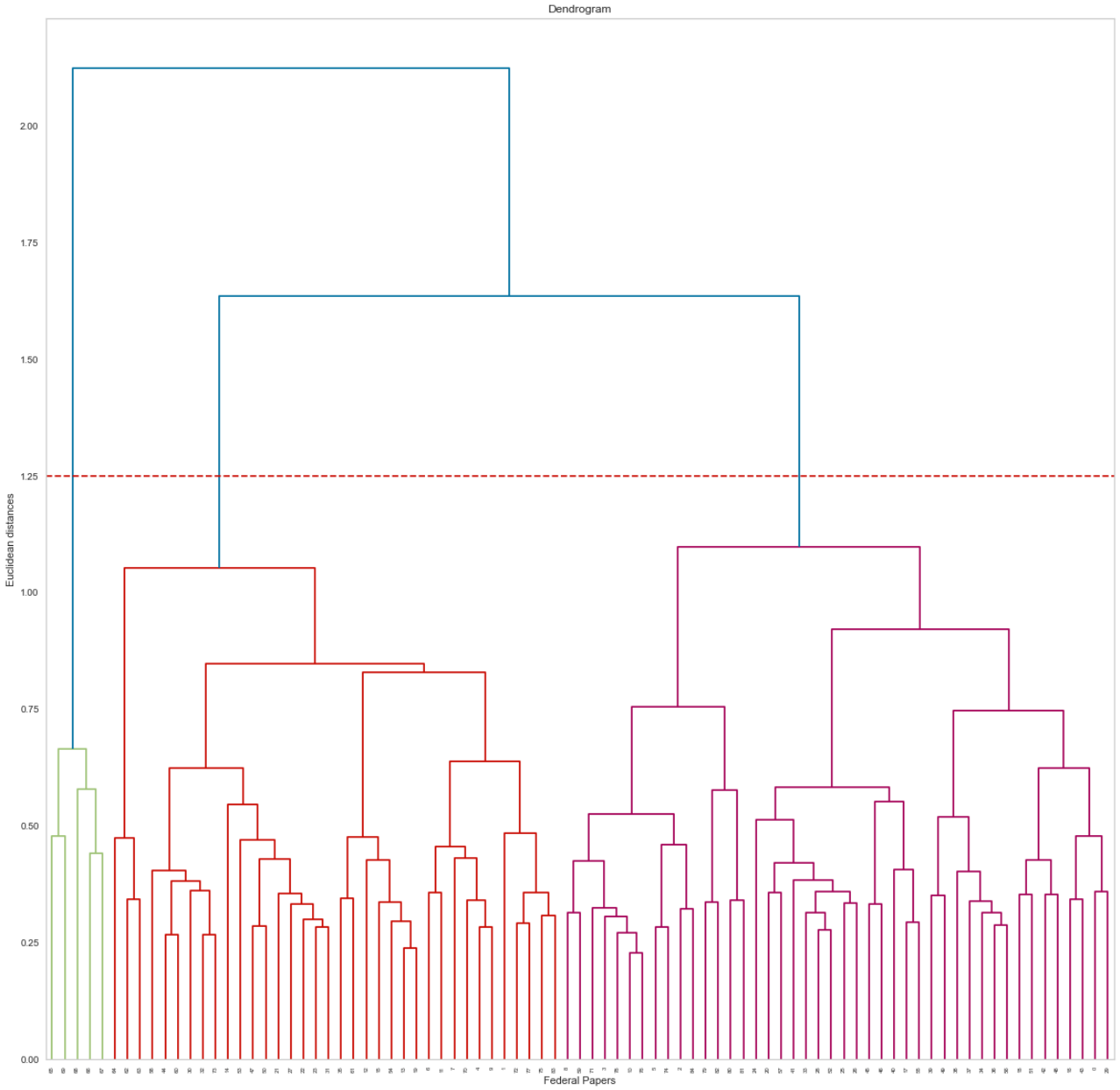
```
skm.f1_score(federalist_papers['author'],federalist_papers['predicted'],average="macro")
```

```
0.5846220527045769
```

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].

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu



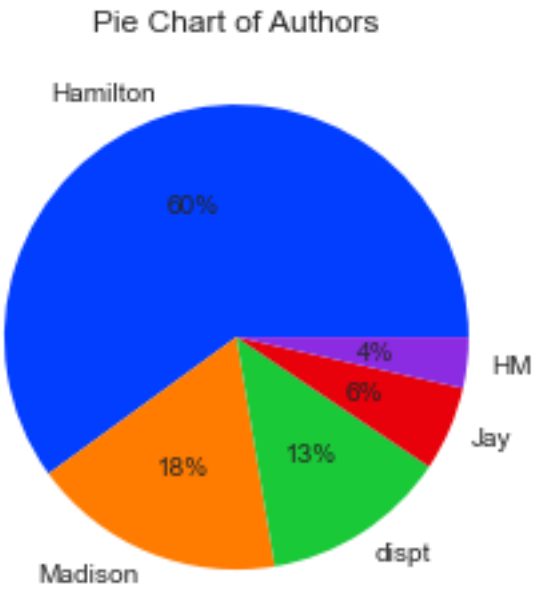
Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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

```
skm.f1_score(federalist_papers['author'],federalist_papers['predicted'],average="macro")
```

```
0.696991150442478
```

CONCLUSION



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.

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

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

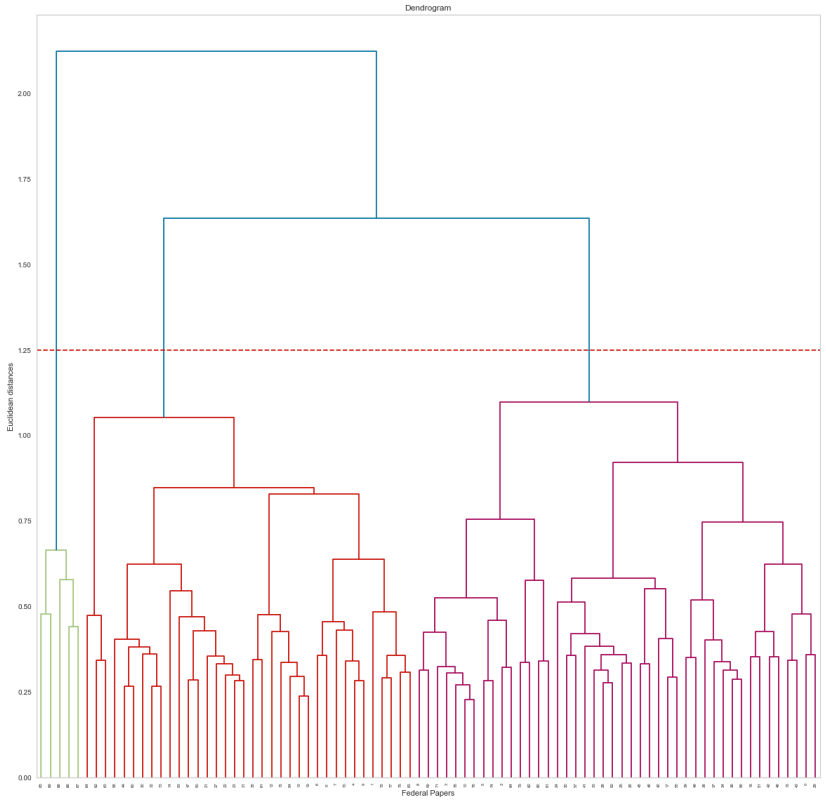
   c. Madison's essays are being grouped in cluster #1, #4

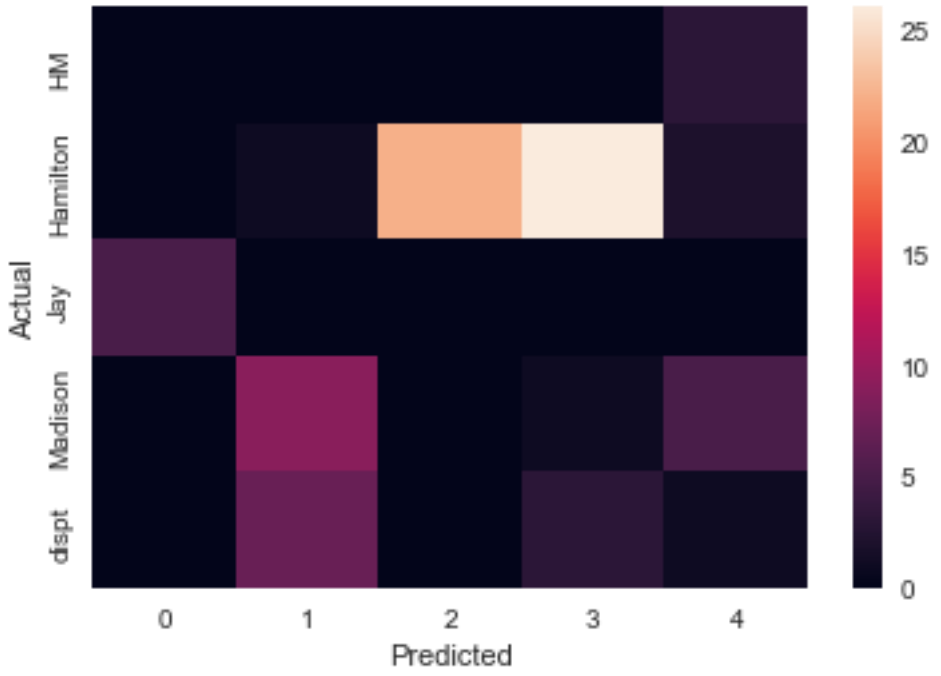
Predicted	0	1	2	3	4
Actual					
HM	0	0	0	0	3
Hamilton	0	1	22	26	2
Jay	5	0	0	0	0
Madison	0	9	0	1	5
dispt	0	7	0	3	1

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

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu





Hierarchical Clustering

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**

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu

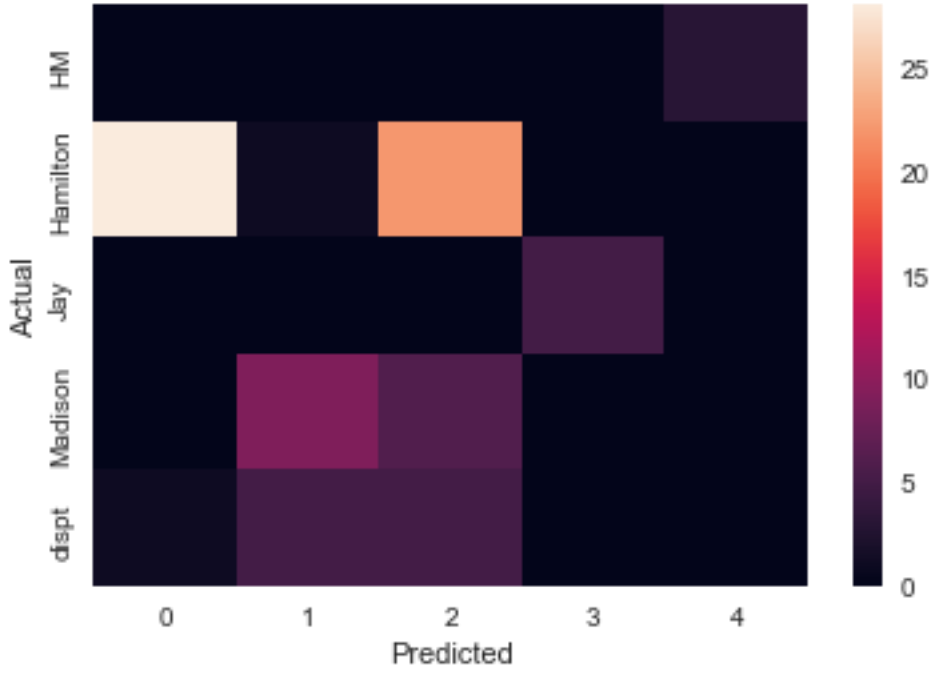
Predicted	0	1	2	3	4
Actual					
HM	0	0	0	0	3
Hamilton	28	1	22	0	0
Jay	0	0	0	5	0
Madison	0	9	6	0	0
dispt	1	5	5	0	0

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

Chaithra Kopparam Cheluvaiah
SUID: 326926205
ckoppara@syr.edu



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.