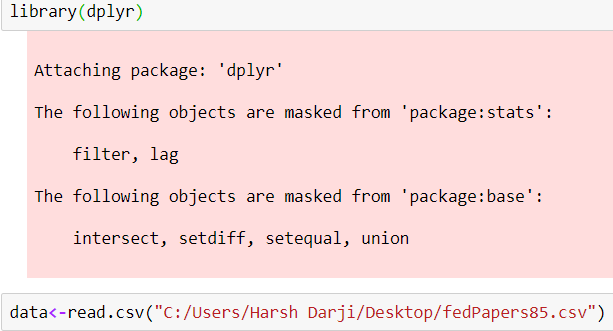
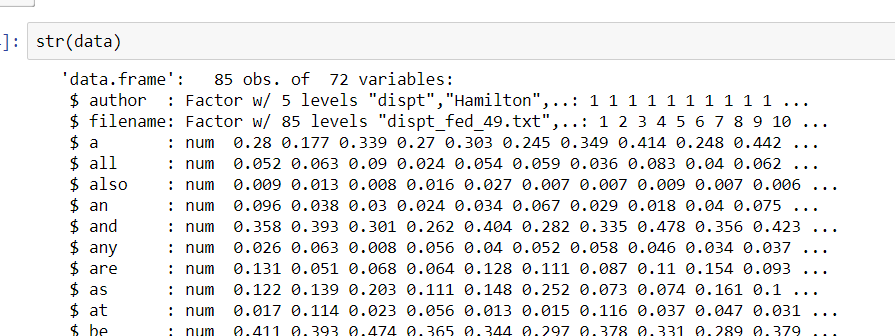
1. **Loading the data**



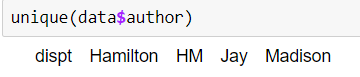
1. **Analyzing the structure**



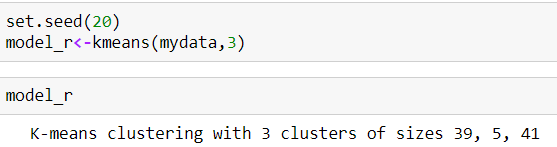
We observe that we have 85 observations for 72 variables and author and filename variables are in factor(nominal) form while rest of them are in numerical form, so we drop them for now and create a new data frame mydata.



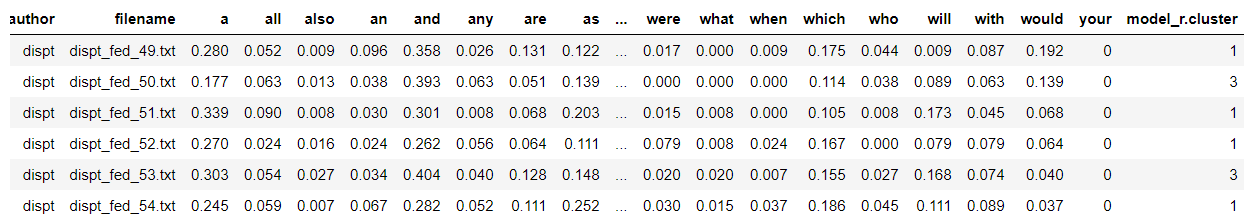
As mentioned in the question we have 3 authors, so we try and identify them, here we disregard disputed papers as author and we choose 3 centroids for our clustering algorithm.

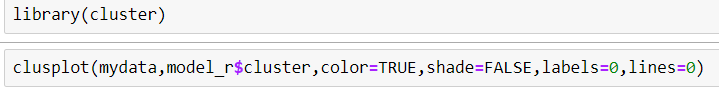


1. **K-Means Clustering**



We obtain 3 clusters with 39, 5, 41 observation in them. Let’s visualize, so that we understand it better.



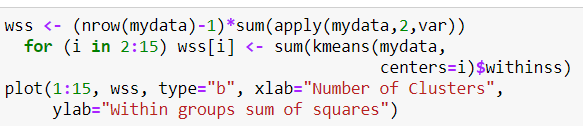


A close up of a map

Description generated with very high confidence

We observe that there is an overlapping between clusters, which implies that we have chosen wrong number of clusters. So, lets use elbow chart to find correct number of centroids/clusters.

**Elbow Chart:**

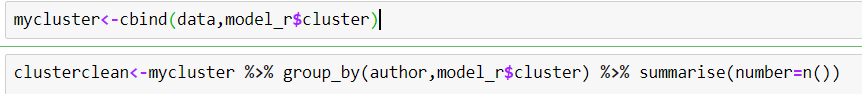


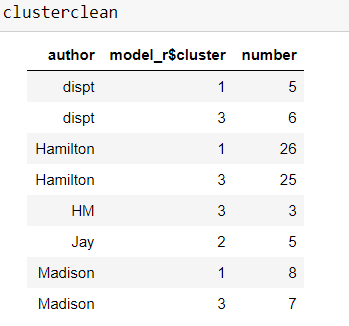
Plot:

A screenshot of a cell phone

Description generated with very high confidence

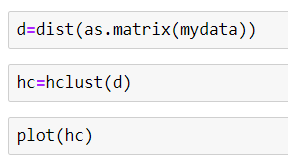
There is no clear point which implies clear clustering where no observation overlaps. So we stick with 3 clusters and group our variables accordingly.





Here, disputed text is clustered into cluster no 1 and cluster no 2, While, all of Jay’s text are clustered into same cluster without overlapping which is a win. In cluster 1, majority of text are of Hamilton followed by Madison and disputed which might suggest that disputed texts are of Hamilton. However, in cluster 3 there are Hamilton’s, Madison’s, and disputed texts are clustered together. But there is no enough evidence that disputed texts might be a collaboration. From, cluster 1 and cluster 3 we can say that Hamilton and Madison might have used similar words for their text.

1. **HAC**

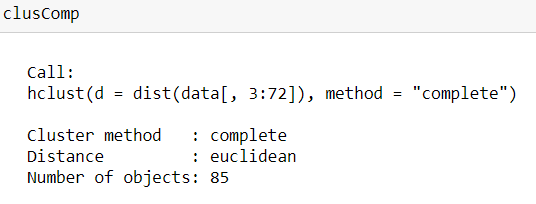


A close up of a logo

Description generated with very high confidence

We use 2 methods, complete and average to observe HAC.





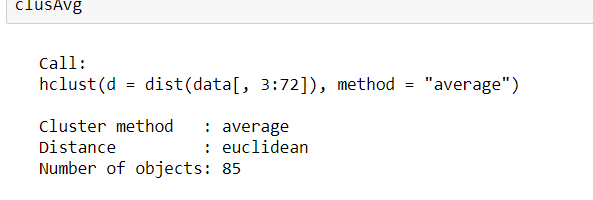
HAC-complete uses Euclidean distance method and cluster method is complete.

A close up of a logo

Description generated with very high confidence

We observe that Jay’s text is clustered separately. While Hamilton and Madison’s text are overlapping. We observe that the co-authored paper HM are placed between Madison’s paper which applies that though papers were co-authored the real writing was done by Madison. Also, out of 11 disputed papers 3 are placed between Madison which suggests Madison have done the writing while 8 are placed between Hamilton.

**HAC-Average:**



Even HAC -Average has used Euclidean distance but the method is average.

A close up of a logo

Description generated with high confidence

Here, Jay’s paper is on left end of a dendrogram which was the case in complete HAC and K-means as well. However, HAC-complete Madison and Hamilton’s text seem to be overlapping unlike HAC-complete which suggests that words used are different. Also, on the right side of dendrogram there are disputed papers between Madison which implies Madison has authorship for those disputed papers.

**Conclusion:**

We conclude that though K-means clustering gave us clear clusters, but it didn’t help enough to derive conclusion except for Jay’s text. However, HAC both complete and Average method gave great insights but it’ s difficult to say who wrote the Disputed papers. Either they could be co-authored, or they might have used each other’s writing.