Forrest Fallon  
HW 4 – Clustering  
IST 707  
10/31/21

This week we will be demonstrating the ability to use sophisticated data mining algorithms to distinguish differences in historical writing styles. Using and tuning these algorithms will hopefully allow us to make some concrete assumptions surrounding the dataset at hand (Federalist Papers), and determine if a specific category of unknown author and determine whether or not they were written by Hamilton or Madison.

**Step 1: Load and Clean the data  
  
Text

Description automatically generated  
**

As seen above, some light cleaning was necessary in order to groom our data for further analysis. The file had each author listed as a column which would cause problems with further clustering; we also assigned the filenames as the row names, in order to distinguish which authors used which words the most.

**Step 2: K means**

****

We will start the K Means process with creating a model with 3 centers; ideally, the data will play nicely with minimal overlap allowing for further progress to be made. Let’s find out:

Diagram

Description automatically generated

Okay! Clearly some overlap going on above, this indicates that there is still some tuning to be done to our K Means model. We will utilize the Elbow Method in order to determine what the ideal number of clusters to use:



Chart, line chart

Description automatically generated

The Elbow Method has shown us that 4 clusters is the ideal amount, so let’s try our clustering again and see what we find:

  
Diagram

Description automatically generated

Clearly there is still some overlap, however, the plot above has some noticeable improvements worth talking about. The 4th cluster seems to have created two appropriate groupings for the left-hand side of the graph, as opposed to one giant cluster in the 1st chart. This indicates that there are similarities worth noting within those two groups. Let’s create a chart to show us where each author lies within the clusters we have created:

  
Text

Description automatically generated  
 Upon inspecting this table, we can begin to see some patterns emerging. The Disputed authors data fits cleanly into clusters 2 and 3; meanwhile Hamilton’s most appearances are within clusters 2 and 3, and Madison’s ONLY appearances are within clusters 2 and 3. This tells us that dispt, Hamilton, and Madison could be considered related somehow. Let’s deploy some HAC analysis to dig deeper.

**Step 3: HAC**

We will start by creating a “Complete” and “Average” model in order to observe HAC.



A picture containing chart

Description automatically generated



A picture containing diagram

Description automatically generated

To the naked eye, the above charts probably seem like the most complicated version of a March Madness bracket. However, upon further inspection, we can make some assumptions based off what these HAC diagrams are telling us.

First, the lines between Hamilton and Madison are clearly blurred. The data heavily favors Hamilton based on the number of appearances within the dataset, and his presence is most definitely known within these two diagrams. The relationship between Madison and Hamilton can be noticed when considering the fact that anywhere Hamilton appears within these charts, Madison is always nearby. This means that the algorithms are having a hard time separating the two, and that their writing styles are very similar.

When taking Dispt into consideration, there are many times where you can find Dispt sandwiched between both Hamilton and Madison. For example, here are four separate instances of these three authors mingling within different clusters:   
Table

Description automatically generatedTable

Description automatically generated with medium confidenceA picture containing table

Description automatically generatedA picture containing text

Description automatically generated

What this tells us is that it is likely that Dispt could be either Hamilton, Madison, or even both! The likely outcome would be that Hamilton and Madison were collaborating frequently through the creation of the Federalist Papers, and perhaps even writing segments for the other in times of need. This would absolutely create similarities and infrequencies in word usage, which creates for some extra analysis to be done on our end for an exercise like this.

In conclusion, data mining is yet another tool in our arsenal to help “tighten the focus” on any information available from a dataset. The picture will not be perfectly clear, although it rarely is, but we can now say for certain that Hamilton and Madison absolutely shared similar writing styles. The disputed papers within this dataset can likely be traced back to either Madison or Hamilton, however, without actually being there to watch them write these papers, we will need some more advanced analysis to say with 100% certainty we know who wrote what.