Nick Videtti

nvidetti@syr.edu

Assignment 4

Clustering

Hierarchical Clustering and Dendrogram Analysis

August 8th, 2022

| Hamilton or Madison? | |
| --- | --- |
| **Introduction** | The Federalist Papers were written by Alexander Hamilton, James Madison, and John Jay, even sometimes by a mix of them. For the most part, the authors of each of these are known, but there is some dispute over who wrote some of them.  In total, there are 85 papers, with 11 of those having a disputed author. The vast majority were written by either Hamilton or Madison, and a smaller amount were written by either Jay or Hamilton and Madison together. The 11 papers will be referred to as the following.    Data have been given that contain some words from each paper. Using those data, the author of the 11 disputed papers can be predicted. For both practical and technical reasons, it will be assumed that the disputed papers were either written by Hamilton or Madison individually, the practical reason being that only 8 of the 74 papers with known authors were not written by either Hamilton or Madison individually. |
| **Analysis** | To answer this question, hierarchical clustering will be performed using complete linkage. This will be done with both Euclidean and Manhattan distance, and the best will be used for conclusions.  First the data must be pulled into R and prepped for the clustering. This includes changing column names, removing the documents that aren’t either exclusively from Hamilton or Madison, and changing the author names to file names for the disputed papers so that author can be used as the labels for the clustering.  Next is to do the clustering. First, Euclidean distance was used, then Manhattan distance. Using the author as the clustering labels allowed for the comparison of each distance measure by looking to see how Hamilton and Madison are clustered. This allows the dendrograms of the clustering using each distance measure to be examined, which gives an idea as to which measure is better to use for the final clustering.  Lastly, making note of each of the disputed papers, the dendrogram of the chosen method will be examined once again. The most common name between Hamilton and Madison that surrounds each disputed paper will be used to predict who wrote the paper. |
| **results** | Euclidean Distance Measure:    Manhattan Distance Measure:    It appears that the Manhattan distance measure created the better clusters, so that will be used to predict each of the disputed papers.   | dispt\_fed\_49.txt | Prediction: **Madison** | | --- | --- | | dispt\_fed\_50.txt | Prediction: **Hamilton** | | dispt\_fed\_51.txt | Prediction: **Madison** | | dispt\_fed\_52.txt | Prediction: **Madison** | | dispt\_fed\_53.txt | Prediction: **Madison** | | dispt\_fed\_54.txt | Prediction: **Madison** | | dispt\_fed\_55.txt | Prediction: **Hamilton** | | dispt\_fed\_56.txt | Prediction: **Madison** | | dispt\_fed\_57.txt | Prediction: **Madison** | | dispt\_fed\_62.txt | Prediction: **Madison** | | dispt\_fed\_63.txt | Prediction: **Madison** | |
| **conclusions** | It seems that many of the papers that are disputed were likely written by James Madison, at least according to the technical findings. This is surprising because Hamilton seemed to have written most of the Federalist Papers, but the margin between the number Hamilton wrote and the number Madison wrote is perhaps much smaller than what is agreed upon.  It was predicted that Hamilton wrote two of the disputed papers and Madison wrote the rest. The predictions per paper were as follows.   | dispt\_fed\_49.txt | Prediction: **Madison** | | --- | --- | | dispt\_fed\_50.txt | Prediction: **Hamilton** | | dispt\_fed\_51.txt | Prediction: **Madison** | | dispt\_fed\_52.txt | Prediction: **Madison** | | dispt\_fed\_53.txt | Prediction: **Madison** | | dispt\_fed\_54.txt | Prediction: **Madison** | | dispt\_fed\_55.txt | Prediction: **Hamilton** | | dispt\_fed\_56.txt | Prediction: **Madison** | | dispt\_fed\_57.txt | Prediction: **Madison** | | dispt\_fed\_62.txt | Prediction: **Madison** | | dispt\_fed\_63.txt | Prediction: **Madison** | |