The Federalist Paper: An Enduring Mystery to True Authentication

HW 4: Clustering

Syracuse University: IST 707

Gavin Grosswald

Introduction

The Federalist Papers, a pivotal piece of American political history, present a true mystery: who authored the disputed essays, Alexander Hamilton or James Madison? Written under the pseudonym "Publius," these essays played a role in advocating for the ratification of the United States Constitution. Despite the contributions of Hamilton, Madison, and John Jay to this work, a subset of eleven essays remains ambiguously labeled "Hamilton or Madison." Data scientists across the globe have tried to solve this mystery using computational methods, building on the groundwork laid by statisticians Mosteller and Wallace in the 1960s. By analyzing the frequency distributions of common function words within the Federalist Papers, modern data mining techniques offer a fresh perspective on who the true Author may be. Using clustering algorithms such as k-Means, PCA, and HAC, researchers have attempted to discern subtle patterns in language usage that could reveal the secrets behind these disputed essays.

Through studying the complete body of work and the linguistic makeup of the Federalist Papers, scientists aim to uncover unique stylistic nuances that can be attributed to the individual Authors. In turn, identifying the true voice of the mysterious 11 essays. Through deep examination and visualization of clustering results, scientists seek to highlight the distinctive writing styles of Hamilton and Madison, while also exploring the possibility of joint authorship scenarios. As technology continues to enhance investigations and analysis, the application of data mining techniques will help to uncover the beautiful historical mystery of the ambiguous Federalist Papers.

Analysis, Models, and Results

About the Data

The dataset comprises the Federalist Papers, a collection of 85 essays advocating for the ratification of the United States Constitution. These essays were initially published between 1787 and 1788 under the pseudonym "Publius," but have since been attributed to Alexander Hamilton, James Madison, and John Jay. The dataset contains 74 essays with identified authors: 51 authored by Hamilton, 15 by Madison, 3 by Hamilton and Madison jointly, and 5 by John Jay, with the remaining 11 essays holding the ambiguous label "Hamilton or Madison." The dataset focuses on the occurrence percentages of function words within the essays, such as "upon" or "for," serving as indicators of writing style. This data forms the foundation for the application of clustering algorithms such as k-Means, PCA, and HAC, offering a computational lens through which to view the mystery surrounding the disputed essays' true authors. Through analysis and visualization of these textual patterns, scientists aim to uncover subtle nuances that may shed light on the identities of the authors behind the Federalist Papers.

Reading / Cleaning the Data

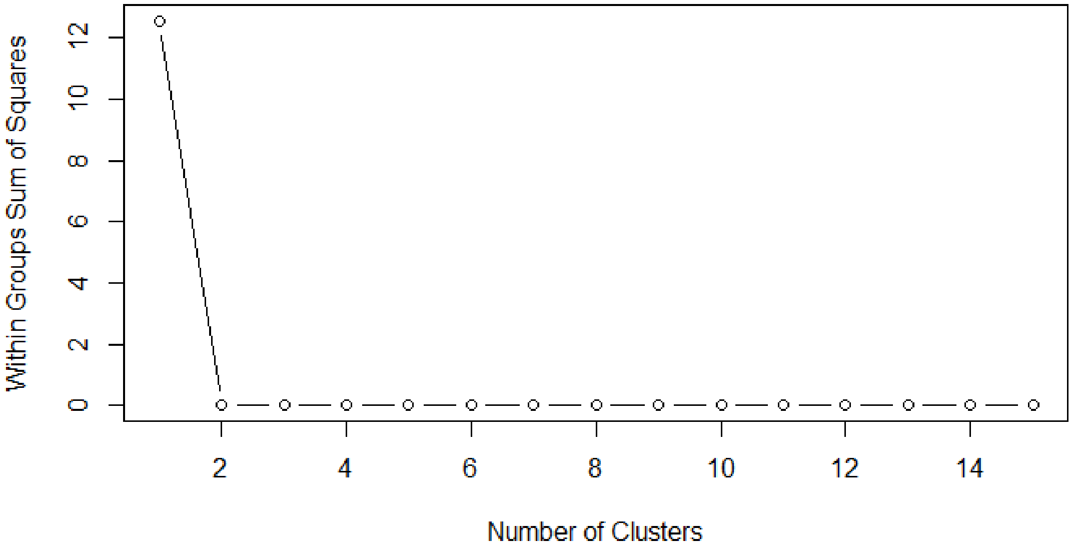
The Federalist Papers csv file is read into R and converted to a dataframe called "FedPapers". In order to best understand the story of the data, functions like dim(), str(), colSums(), summary(), and colnames() are queried. Initial observations are as follows:

- 85 observations of 72 variables
- Columns [,1:2] are <chr> and columns [,3:72] are <num>
- Columns [,1:2] relate to the author and essay file and columns [,3:72] represent feature words found among all 85 essays
- The values are the percentage of the word occurrence in the specific essay (i.e., if the word "upon" appeared 3 times, and the total number of words in the essay is 1,000, the value is 3/1,000 or 0.3%)

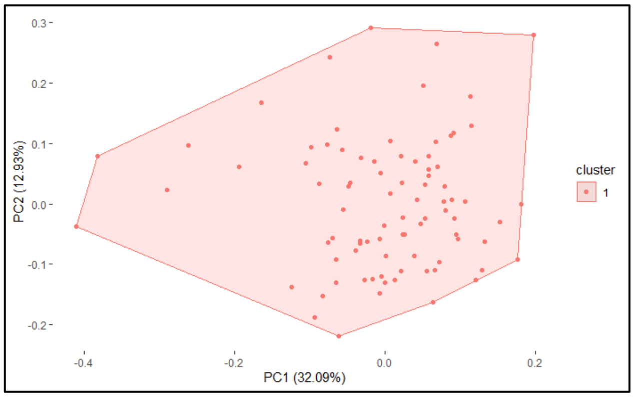
To prepare the data for cluster analysis, columns [,1:2], "author" and "filename", are removed which creates a completely numeric dataset.

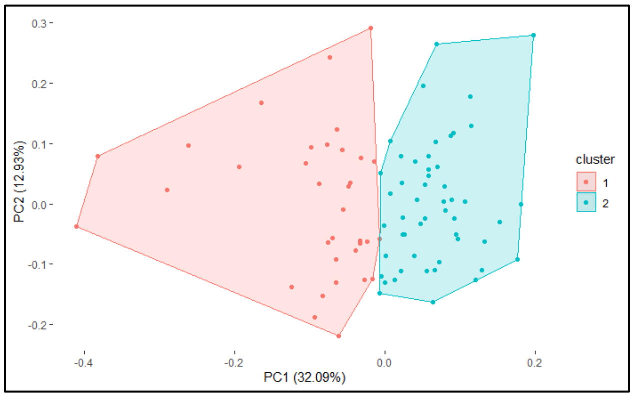
Data Analysis

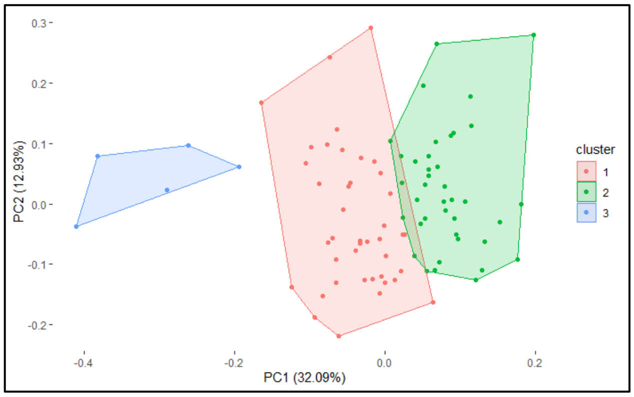
The first step to cluster analysis is to identify the optimal number of clusters. The Within-Cluster Sum of Squares (WSS) function returns the following chart and the point at which the chart flatlines is considered the optimal number of clusters. As seen below, that value is two.

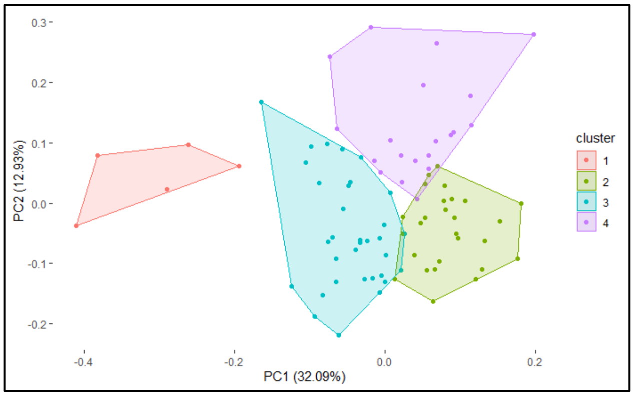


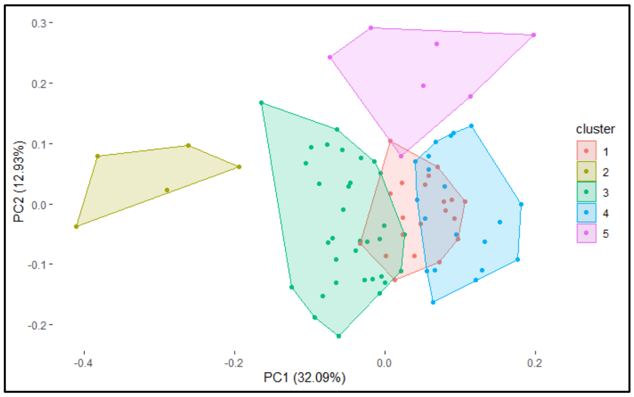
Following this determination, k-means, which is a widely used unsupervised machine learning algorithm designed for clustering data into distinct groups based on similarity, is leveraged. The goal of k-means clustering is to partition a dataset into k clusters, where each observation belongs to the cluster with the nearest mean, or centroid. For sake of completeness, the k-means analysis is run using a k value equal to [1:6] and the below cluster plots are generated.

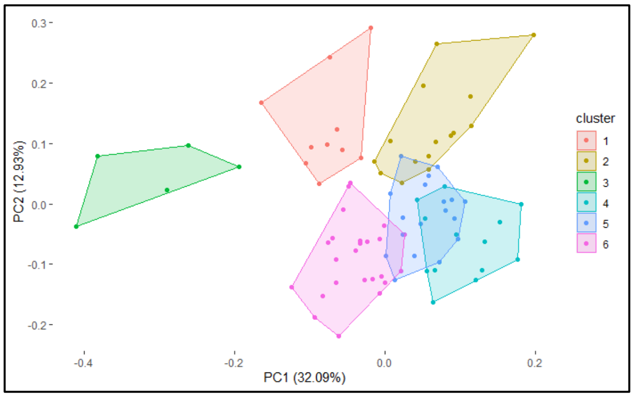








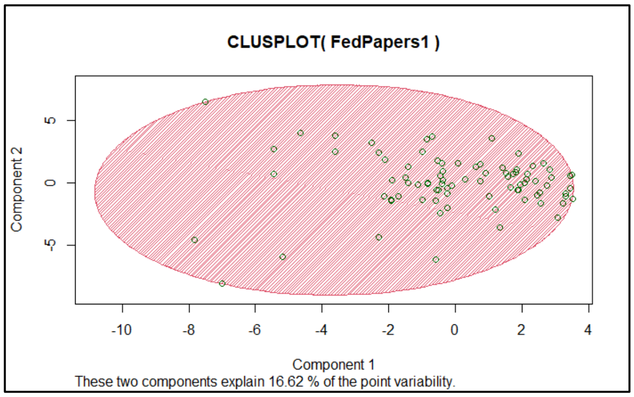


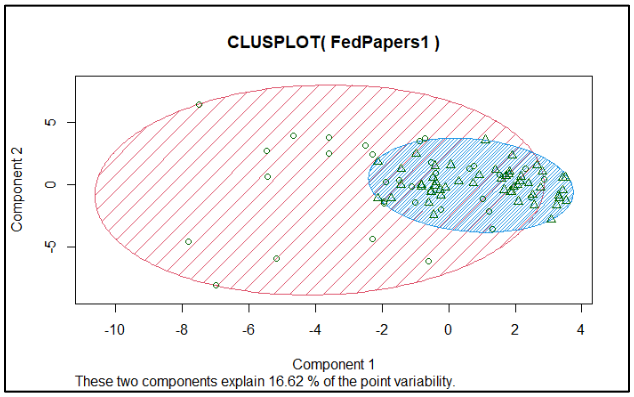


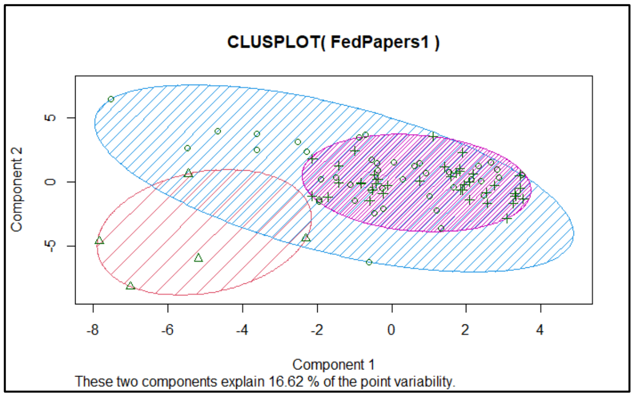
The k-means charts provide useful information and initial observations are found below:

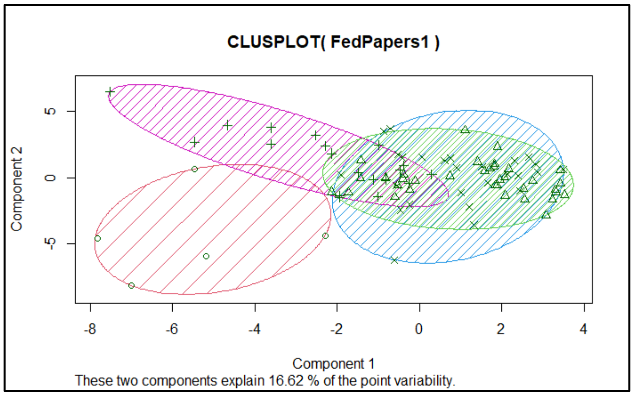
- Cluster 1 can be ignored, included for completeness
- Cluster 2 is successfully plotted as there is no overlap between the two groupings
- Clusters 3 and 4 are relatively successful with only a small amount of overlap
- Clusters 5 and 6 show a good amount of overlap rendering them unsuccessful

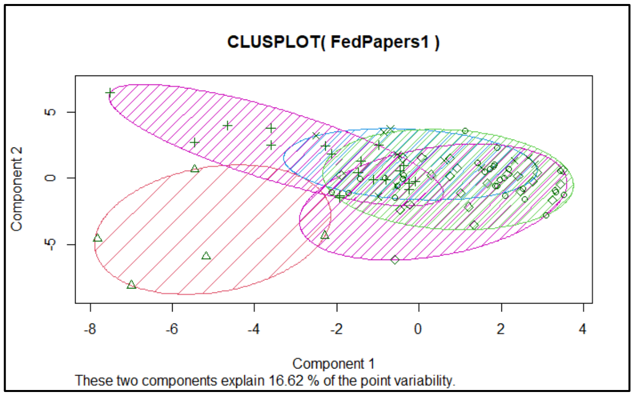
Next, Principal Components Analysis (PCA) is used to help visualize the k-means model. The combination of PCA and k-means allows for more efficient clustering of high-dimensional datasets while preserving important patterns in the data.

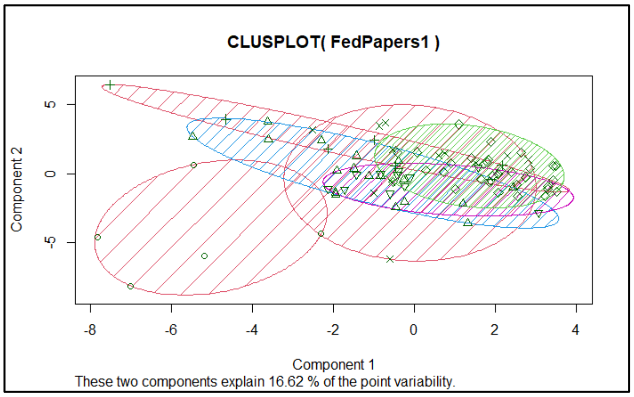




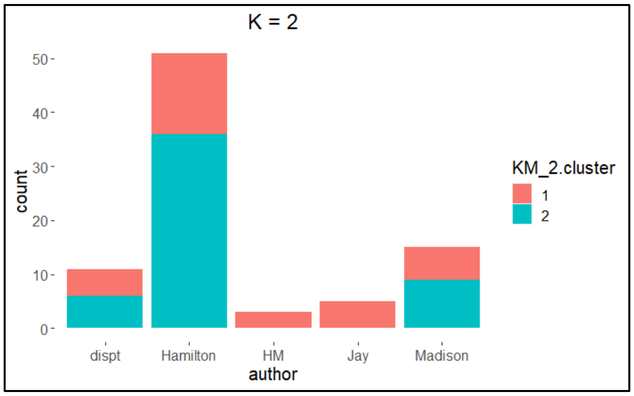


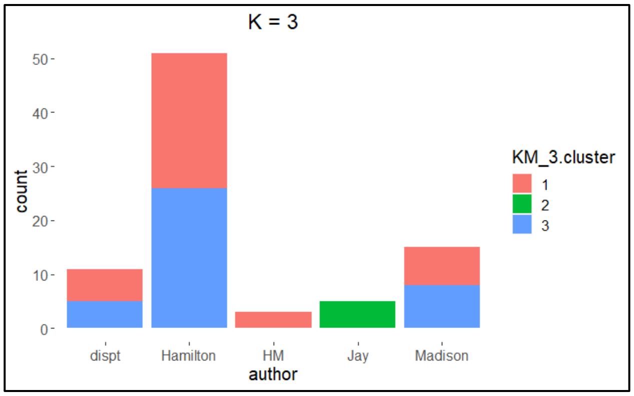


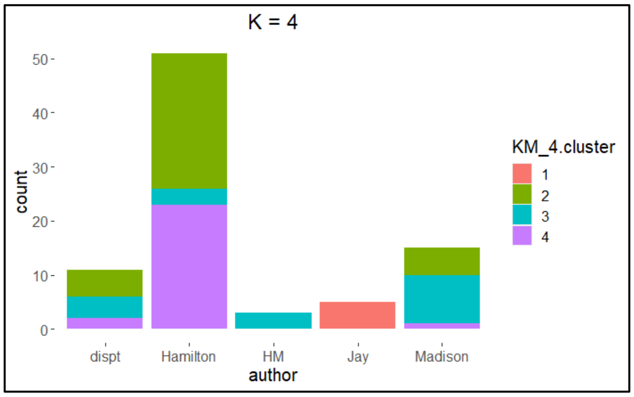




After reviewing the clusters queried as part of the k-means and PCA analysis, it is determined that it would be best to further analyze the data using clusters [2:4]. The clustered groups are then combined with the original “FedPapers” dataframe and inserted into a new dataframe called “FedPapers_clusterDF[2:4]”. This dataframe includes all 85 essays and the cluster that they associate with. To visualize how each author relates to each cluster, the results are plotted on a bar chart using k=2, k=3, and k=4.







Although a two-cluster approach is the most successful as represented by the WSS function and k-means cluster analysis, it leaves questions as to who the identity of the 11 papers in question may be. The four-cluster bar chart provides meaningful insight into the individual authors and their unique writing tendencies. Observations are included below:

- Jay is isolated to cluster 1 meaning his unique writing style differs from that of the other authors'
- Madison's writing technique includes focus words heavily concentrated in cluster 3 while Hamilton's tendencies reside within both clusters 2 and 4
- The three essays authored by both Madison and Hamilton jointly appear to be heavily influenced by the writing style of Madison as seen by the collection of essays being isolated to cluster 3, the cluster most unique to Madison
- The 11 disputed essays contain attributes from each of clusters 1, 2, 4 and it is difficult to make a confident prediction into who the sole author may have been
- Based on the cluster analysis and the bar charts, it is most likely that the 11 disputed essays have influence and were authored by both Madison and Hamilton jointly

Author’s Analysis Notes

The results of the analysis remain ambiguous and below the report’s Author notes reasons why uncertainty into the identity of the Author of the 11 essays exists post-completion of the clustering analysis.

- The dataset includes too many words to be analyzed (85 essays of 70 focus words). There are nearly as many words being analyzed as there are essays. It would be helpful going forward to trim down the number of words and only focus on a subset of 10-15 focus words
- The focus words are very generic such as “as”, “even”, “well”, etc. which are likely to be used by each Author. It would be beneficial to have a subset of more unique words found in the collection of essays to help identify the words more unique to one Author
- K-means algorithm provides random results each time it is run which impacts the data. It is crucial to set the randomization seed to eliminate randomization and create stable, consistent results

Conclusion

The search to uncover the authors of the disputed Federalist Paper’s essay is a mystery that continues to leave data scientists perplexed. Scientists around the world continue to leverage advanced data mining techniques to analyze language patterns in hopes of distinguishing between Alexander Hamilton's and James Madison's unique writing styles. However, despite these efforts, the mystery behind the eleven ambiguous essays remains, highlighting the complexity of language and the need for continued study. With the constant advancement and investment in the data science field, analysis techniques will continue to improve, ultimately helping to solve this mystery once and for all.