**Disputed Federalist Papers – Continuing the Analysis**

This report is a continuation of the report for Homework 4 that focused on the federalist papers and analytical approaches such as K-means clustering and Hierarchical Agglomerative Clustering techniques. This continuation of the Federalist papers studies will include a new model to help distinguish the author of the disputed papers: Decision Trees.

The Federalist Papers, a collection of influential essays published during the late 18th century, played a pivotal role in persuading New York citizens to support the ratification of the United States Constitution. Authored by Alexander Hamilton, James Madison, and John Jay under the pseudonym "Publius," these eighty-five essays initially appeared anonymously in New York newspapers between 1787 and 1788. While the first published edition in 1788 did not reveal the true authors, it wasn't until the 1818 edition by printer Jacob Gideon that the individual contributors were formally identified. Today, The Federalist Papers are regarded as an invaluable resource for comprehending the Constitution's original intent.

Machine learning clustering methods have emerged as powerful tools to explore and extract patterns from vast datasets, making them highly applicable to historical authorship attribution of The Federalist Papers. By leveraging these techniques, researchers and historians can seek to determine the authors behind each essay based on linguistic patterns, writing style, and content. One widely used clustering method in this context is K-Means, which groups documents into distinct clusters based on their similarity, potentially revealing underlying authorship patterns among the essays.

Unraveling the true authors of The Federalist Papers through machine learning clustering methods not only adds to historical knowledge but also showcases the adaptability and power of artificial intelligence in the analysis of complex literary works. By providing insights into the distinct writing styles and ideological influences of Hamilton, Madison, and Jay, this approach sheds light on the minds behind one of the most influential documents in American history, reaffirming the significance of The Federalist Papers as a timeless testament to the formation of the United States' constitutional framework.

**Analysis and Models**

This section will provide additional information in the data that was used to complete this study, the data preparation & cleaning process, and an overview of the model used to conduct additional analysis. For this assignment, the team will primarily utilize K-means clustering and Hierarchical Agglomerative Clustering (HAC) techniques.

K-means clustering is a fundamental unsupervised machine learning technique used for grouping data points into distinct clusters based on their similarity. The algorithm aims to partition a given dataset into K clusters, where each cluster is represented by a centroid, typically the mean of the data points within that cluster. The process involves iteratively assigning data points to the nearest centroid and then updating the centroids based on the newly assigned points. This iterative process continues until convergence, minimizing the sum of squared distances between data points and their respective centroids. K-means is widely used in various fields, such as data analysis, image segmentation, and customer segmentation, to reveal underlying patterns, structure, and relationships within data. However, it requires an initial guess for the number of clusters (K) and can be sensitive to the initial placement of centroids, which can impact the quality of the resulting clusters.

Hierarchical Agglomerative Clustering (HAC) is a widely used technique in the field of data analysis and machine learning, specifically in the realm of unsupervised clustering. HAC operates by iteratively merging individual data points or existing clusters into larger clusters based on their similarity, gradually forming a hierarchical tree-like structure known as a dendrogram. The process begins with each data point as a separate cluster and, in each iteration, the two closest clusters are combined, effectively reducing the total number of clusters. This continues until all data points are part of a single, comprehensive cluster or until a predetermined number of clusters is reached. HAC's strength lies in its ability to reveal both global and local structure in data, accommodating various distance metrics and linkage methods to measure similarity between clusters. The resulting dendrogram provides insights into the data's hierarchy, aiding in interpretation and decision-making for further analysis.

**About the Data & Business Scenario:**

In the process of applying machine learning clustering methods, the first step is to transform the text data into a suitable format for analysis. This typically involves preprocessing the essays, removing irrelevant information, and representing the remaining textual data numerically to create a feature matrix. Next, researchers can perform dimensionality reduction to focus on the essential signal while mitigating noise that might arise from varying writing styles across the essays. Once the data is appropriately prepared, K-Means clustering is employed to identify clusters of essays that share similar characteristics, potentially grouping together the works of the individual authors.

*Figure 1 – Portion of the loaded csv file.*

A table of numbers and letters

Description automatically generated

“author": This column indicates the author of the Federalist Paper section.

"filename": This column specifies the filename or identifier associated with the Federalist Papers. In this case, the paper in question is titled "dispt\_fed\_49.txt."

Numeric values: The subsequent columns represent the frequency or occurrence of various words within the text of the Federalist Paper. Each numeric value corresponds to the frequency of a specific word within the paper.

For example, looking at the first few numeric values:

* 0.280: Indicates the frequency of the first word, ‘a’ in the list within the text of "dispt\_fed\_49.txt” file.
* 0.052: Indicates the frequency of the second word, ‘all’ within the “dispt\_fed\_49.txt” file
* 0.009: Indicates the frequency of the third word, “also” within the same text file.
* This pattern occurs for every word within the entire data frame.

This representation of data is part of a larger dataset that includes multiple Federalist Papers, each with its corresponding author, filename, and word frequency information. Researchers and analysts can use this data to study patterns, writing styles, and linguistic features within the Federalist Papers, potentially aiding in authorship attribution or gaining insights into the authors' preferences and writing techniques to help identify who the authors are for the disputed texts.

**Null or Missing Values:**

There are no missing values in this dataset as shown in the figure below:

*Figure 2 – Confirmation of no NULL values in each of the columns*

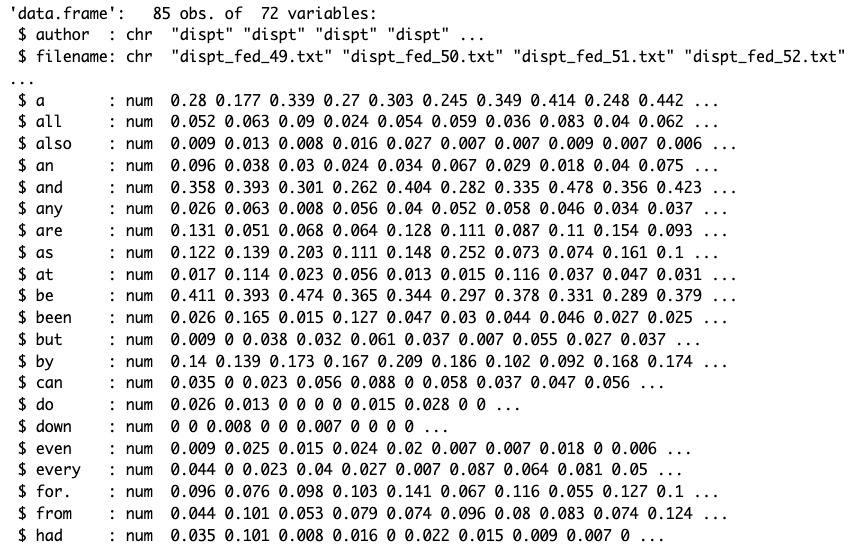
**A close-up of a computer screen

Description automatically generated**

**Author Breakdown:**

Before conducting any form of analysis, the team wanted to understand the author breakdown for the federalist papers. Based off the data provided to the team, we can see that will find 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. The remaining 11 essays, however, is authored by “Hamilton or Madison”. The rows of this data set are words that can be found in all of the documents within the Federalist Papers collection. Based on the output, we can see that there are 85 words (rows).

*Figure 2 – Federalist Papers dataset structure*

**

To help illuminate the author breakdown, the team created two forms of visualizations to help highlight the number of texts each author contributed to the Federalist papers collection.

*Figure 3 – R Output for Author Breakdown.*

A screenshot of a computer

Description automatically generated

To help visualize on the author breakdown, the team created a pie chart to help illuminate which author contributed the most to the Federalist Papers:

*Figure 4 – Pie Chart Author Breakdown*

A pie chart with different colored sections

Description automatically generated

**Sub-setting Data by Author:**

In order to answer the question of who authored the disputed texts of the Federalist papers, it is important to create data frames for each author. To do that, the team created a simple command that would filter on the author column for one of the 4 known authors:

*Figure 6 – R Code to Create filtered Data frames per author*

A screen shot of a computer

Description automatically generated

**Calculate Column Means**

The team then decided it needed to calculate the average of the values in the author data frame. To accomplish this, the team created a R-code function to help calculate the means of a column. This function takes a matrix object as input, calculates the column means for specific columns in that matrix, and creates a new vector containing those column means. The resulting vector is then printed. Note that the specific columns for which the means are calculated are determined by starting from the third column (index 3) up to the last column in the input matrix. Each one of the author data frames will be passed into the function.

*Figure 7 – CreateWordMean Function Signature*

A white screen with green text

Description automatically generated

**Additional Function Explanation**

* y <- ncol(x): This line calculates the number of columns in the input matrix x and stores it in the variable y. This value will be used to determine the range of columns to calculate column means for.
* x <- colMeans(x[,3:y]): Here, the code calculates the column means for columns starting from the third column (index 3) to the last column (y). The colMeans function computes the mean value of each column in the specified range and assigns the resulting vector of means to the variable x.
* newVector1 <- c(): An empty vector named newVector1 is created. This vector will be used to store the calculated column means.
* for (i in 1:length(x)) { newVector1[i] <- x[[i]] }: This loop iterates over the indices of the vector x (which contains the calculated column means) and assigns each mean value to the corresponding position in the newVector1 vector.
* print(newVector1): Finally, the code prints the newVector1 vector, which now contains the calculated column means.

**K-Means Clustering**:

For this study, the team focused on K-means clustering and HAC techniques to draw insights. This portion will highlight the process the team took to conduct K-Means Clustering analysis.

First, the team created a new data frame that did not include the author column, but used the remaining column values from the raw data file. Next, the team re-structured the data frame to have the filenames become the row values. The team tried to find the optimal number of clusters to be used later. Please reference the output below for optimal clusters:

*Figure 8 – Optimal Clusters Visual – WSS Method*

*A graph with a line

Description automatically generated*

After generating both plots, there was no clear “elbow” in the WCSS plot to highlight that there is a cluster amount to use. Since there was no clear elbow in the first plot, the team made a general guess on what the optimal clusters was to use, which was 5.

*Figure 10 – Optimal Number of Clusters – Silhouette Method*

A graph with a line and a number of clusters

Description automatically generated

Based off this output, we can see that there is a peak at 2 clusters.

**Results**

Based on the output above, the team ran two trials to see which cluster amount, 2 or 5 would produce the desired results to help draw insights on who authored the disputed papers. Here are the results below.

*Figure 11 – Cluster Plot – 5 Clusters*

*A diagram of a graph

Description automatically generated*

A graph with different colored squares

Description automatically generated

When using the “elbow method”, the team was not able to visually see a distinct bend in the graphic as per Figure 8 above. This forced the team to pick a value based on instinct versus anything concrete. The value the team chose was 5 and produced the above visuals. What the team noticed is that this amount of clusters mirrors the behavior of the author breakdown of the Federalist Papers. What came out as interesting to the team was that cluster number 2 annoted as the brown segment on the second plot of Figure 11, it suggests that there similar writing styles or word usage between two authors: Hamilton and Madison. This is demonstrated in the visual as that it spans four of the 5 possible author buckets. Just by going at face value, we can see that Hamilton and Madison are the two prominent authors that may have authored the disputed papers. However, the team was not able to derive concrete numerical data to support this claim. Therefore, the authorship of the disputed papers is still left unanswered.

*Figure 12 – Cluster Plot – 2 Clusters*

A diagram of a diagram

Description automatically generated

A graph with red and blue squares

Description automatically generated

When using the value of 2 which was the team’s silhouette output, the results were drastically different. The algorithm over grouped the textual documents in a way that it was impossible to differentiate which author wrote what. Furthermore, this cluster value produced results that was not similar or mirrors the same behavior of the factual author breakdown that was annotated in figures 3 and 4. Lastly, the output utilizing this value did not provide clarity, but only created more questions. Based on this, the team decided that the cluster value of around 5 may be the optimal cluster value.

**HAC – Dendrograms**

The purpose of Hierarchical Agglomerative Clustering (HAC) and cluster dendrograms is to reveal and visualize the inherent structure or patterns within a dataset by grouping similar data points together into clusters. HAC is a popular technique in the field of unsupervised machine learning and data analysis, while cluster dendrograms are graphical representations that help interpret the results of HAC.

HAC is a clustering algorithm that builds a hierarchy of clusters by successively merging or "agglomerating" data points or existing clusters based on their similarity. It starts with each data point as its own cluster and then iteratively combines the closest clusters until all data points are part of a single comprehensive cluster or a predetermined number of clusters is reached. HAC provides a natural way to explore both global and local structure in the data and allows for the identification of nested clusters, which can be especially useful when the number of clusters is not known beforehand.

Cluster dendrograms are tree-like diagrams that visually represent the results of HAC. They showcase the hierarchical relationships between data points or clusters, providing a way to understand the grouping structure in the data. In a dendrogram, each leaf node represents an individual data point, and as you move up the tree, clusters are formed through agglomeration. The height at which clusters are merged on the dendrogram corresponds to the level of similarity or dissimilarity between them. Longer branches indicate greater dissimilarity, while shorter branches signify higher similarity. By "cutting" the dendrogram at a certain height, you can determine the number of clusters you want to form.

*Figure 13 – Cluster Dendrograms with K = 5*

A diagram of a city

Description automatically generated

**A diagram of a cluster

Description automatically generated**

*Figure 14 – Cluster Dendrograms with K = 2*

**A diagram of a city

Description automatically generated**

**A diagram of a cluster of buildings

Description automatically generated**

The results from the two HAC trials mirrored the results from our K-mean’s clustering experiments. When we used a higher cluster value, we were able to see results that were able to segment texts into author groups more clearly. When we used a cluster value of 2, the grouping was more inclusive and made it very difficult to draw any conclusion. This makes sense that if we are only creating 2 buckets, of course the algorithm will not segment the authors in a precise manner. Ultimately, the salient lesson the team learned from this study is that a cluster amount should be around the same amount of categories in your dataset. For this example, we have five author pools, so the amount of nodes for clustering techniques should be around that number.

**Decision Trees** -

Decision trees are powerful and interpretable machine learning models widely used in identifying patterns in writing styles for authorship attribution. They work by recursively partitioning the data into subsets based on the most discriminative features, making them particularly useful for analyzing textual data. In the context of identifying the author of the disputed Federalist Papers, decision trees can examine the frequency and distribution of specific words or phrases that are characteristic of an author's writing style. By selecting the most informative features, decision trees can efficiently distinguish between different authors and classify new texts based on their similarity to known writing patterns. The use of decision trees allows researchers to gain insights into the decision-making process of the model, helping to uncover the distinctive linguistic markers of individual authors.

The initial analysis utilizing various models yielded a significant discovery. By applying the k-means algorithm with 5 clusters, it effectively differentiated papers authored by Jay from those written by Hamilton and Madison. Jay's papers were predominantly grouped within one cluster, while Hamilton's and Madison's works were found in separate clusters. Nevertheless, there were instances of overlap, where joint works of Hamilton, Madison, and their individual papers were clustered together.

To facilitate comparison among different clustering results, a data frame was created, with rows representing individual papers and columns indicating clustering outcomes from diverse models. This arrangement facilitated easier comparisons and in-depth analysis of the results.

For the implementation of decision trees and random forests, a 60% data split was used for training and testing, and predictions on disputed papers were made after retraining the models with all the training data. Both pruned and unpruned decision trees indicated that Madison authored all 11 disputed papers, corroborating the predictions from hierarchical clustering. Internal cross-validation suggested that pruning the tree at a depth of 1 node (excluding the root) achieved the best outcomes.

Random forests exhibited promising results, predominantly predicting Madison as the author for most papers, except for the 7th disputed paper. The random forest's predictions closely aligned with the clustering approach, providing compelling evidence that different models can lead to similar conclusions.

Remarkably, decision trees and random forests provided explicit reasons for their classifications, with the word "upon" emerging as a critical factor distinguishing Madison from Hamilton. The visual representation of the random forest further supported this observation.

Overall, the findings indicate that hierarchical clustering and random forests offer valuable insights and accurate predictions for authorship attribution. The use of diverse models and methodologies enhances our understanding and confidence in the conclusions drawn from the analysis. Additionally, it is important to note that not a single model will be able to accurately classify behavior, but should be seen as a way to confirm or deny results from another implemented model.

*Figure 15 – Decision Tree 1– Author as a function of all attributes*

*A computer screen shot of a code

Description automatically generated*

This output tells you that the tree was constructed using the "not upon" variable to predict the "author" classification. The initial prediction wasn't perfect, with about 21.43% of observations being misclassified at the root node. The tree has more information about how it splits and the error rates at different nodes. The CP values in the table provide insights into potential pruning points to make the tree simpler and less likely to overfit the training data.

A diagram of a group of people

Description automatically generated

A second tree was created using the “upon” variable to predict an author. The initial prediction wasn't perfect, with around 21.43% of observations being misclassified at the root node. The tree's structure and error rates at different nodes are provided, and the CP values help identify possible points where the tree could be pruned for better generalization to new data. Please see the results below:

*Figure 16 – Decision Tree 2*

A white screen with black text

Description automatically generated

A diagram of a graph

Description automatically generated

**Classifying Authors by Text**

The team then used the tree to make predictions on who an author was for a given disputed document. Here is the code and output below:

*Figure 17 – Decision Tree Predictions*

A screenshot of a computer program

Description automatically generated

The code above builds a decision tree model to predict the "author" of a document based on its features. It then uses this model to predict the authors of some documents and provides the predicted author labels as output. The model structure is retained for further use.

A salient contrast between decision trees, random forests, and clustering lies in their ability to provide clear explanations for their classifications. In the case of decision trees, they make decisions based on the frequency of the word "upon," effectively distinguishing between Madison and Hamilton. This same pattern is echoed in the random forest analysis through a visualization named "Distribution of minimal depth and its mean." This visualization highlights that "upon" consistently emerges as the top choice for the root node. The subsequent significant words are "on" and "and." Additionally, the "Multi-way importance plot" visualization underscores the pivotal role of "upon" as the most influential word, closely followed by "on," "and," and "to."

**Conclusions**

In conclusion, our exploration into unraveling the authors of The Federalist Papers via machine learning clustering methods has yielded intriguing insights. While these methods hold promise in shedding light on the potential authors behind the essays, the results obtained from our team's comprehensive experiments have proven inconclusive. The application of techniques such as K-Means has offered glimpses into patterns within the essays, revealing potential authorship cues. However, the results forced the team to speculate and infer versus relying on concrete data to draw conclusions.

The intricacies of definitively determining the authors behind the disputed papers beckon for a deeper examination. To achieve conclusive outcomes, our endeavor requires a multidimensional approach that encompasses not only the application of advanced machine learning techniques but also a nuanced understanding of the linguistic subtleties, contextual cues, and the broader historical landscape of the late 18th century. As the realm of artificial intelligence intersects with historical scholarship, the pursuit of unveiling the true identities behind these influential essays becomes an intricate dance between data-driven exploration and domain expertise.

In the broader context of historical understanding, the team’s ongoing journey underscores the harmonious fusion of machine learning prowess and human interpretative skills. The quest for authorship attribution of The Federalist Papers is a testament to the evolving synergy between data-driven methodologies and nuanced analysis, offering a unique vantage point into the minds of Alexander Hamilton, James Madison, and John Jay. While the journey towards conclusive outcomes continues, our collective effort contributes to a deeper comprehension of this pivotal collection, reaffirming its enduring significance in shaping the foundational principles of the United States' constitutional framework.