Nicholas Lichtsinn

IST 707

Homework 4

**Introduction**

The Federalist Papers were a collection of eighty-five essays pushing New Yorkers to ratify the newly formed United States Constitution. The pieces were first published anonymously in New York newspapers in 1787 and 1788 under the pen name “Publius,” and were written by Alexander Hamilton, James Madison, and John Jay. The Federalist Papers are regarded as one of the most important sources for analyzing and comprehending the Constitution’s original intent.

Seventy-four of the eighty-five articles have an identified author, however there are eleven articles that have a disputed author. Both Alexander Hamilton and James Madison claimed to have written these documents and the truth is unknown. This had led to the use of a variety of machine learning and natural language processing techniques to determine the correct author of the disputed articles.

To figure out the appropriate author for the eleven disputed documents this paper will utilize K-Means Clustering and Hierarchical Algorithm Clustering techniques.

**Analysis**

Data Preprocessing:

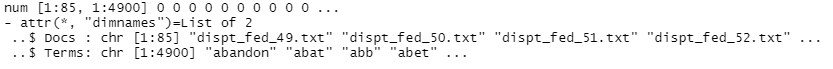
The data files consisted of eighty-five text files, one for each of the Federalist Papers. Showing the author and the written body of each specific article. For the eleven disputed articles the author is unknown.

The documents were then read into a Corpus and the files were stripped of anypunctuation, numbers and spaces to identify word frequencies and create a document term matrix.

Text

Description automatically generated

The word vectors were then normalized to allow the use of standard analysis techniques.



Text

Description automatically generatedAfter normalizing the data was converted to a matrix and a dataframe to make it easier to work with.

A close-up of a document

Description automatically generated with low confidenceA word cloud was then created from the document term matrix of the most frequently used words in the corpus.

This gives an insight into what the texts are concerning and what they are talking about. Mainly people, government, senate etc.

Since the data has been vectorized analyses could be performed and data clusters were created using HAC and K-Means methods. Using the HAC method clusters were created with Euclidean, Manhattan, and Cosine methods as well as a normalized Cosine method as well. From these dendograms were plotted to show the results. For K-Means clustering clusters of two and seven were used as well as Manhattan and Euclidean visualization models.

**Results**

Here are the resulting dendograms from the HAC models after computing a variety of distance matrices:

**A picture containing diagram

Description automatically generated**Diagram

Description automatically generated**Histogram

Description automatically generatedDiagram

Description automatically generated**

From these it appears that the Manhattan and Euclidean models have outperformed the Cosine models as they have all the disputed articles in closely related branches, and they are linked more closely with James Madison.

Next K-Means Clustering was used to see which author the articles related more closely to using a cluster of two and seven.

Text

Description automatically generated

Here are the results of the k-means visualizations for Manhattan and Euclidean methods as they were the highest performing with the HAC models.

Chart, line chart

Description automatically generated

**Chart, line chart

Description automatically generated**

**Chart, bar chart, histogram

Description automatically generated**These results are much harder to see so included is the two cluster visualization and results matrix:

**Table

Description automatically generated**Here the breakdown is shown with all of Madison’s articles and the disputed articles in cluster one and most Hamilton’s articles in cluster two.

**Conclusion**

From the results of the HAC and K-Means clustering models used, James Madison looks to be the author of the disputed Federalist Papers articles. Both models showed the same result and classification.

Historically this does present an issue as there are two of America’s founding fathers claiming authorship of key documents in out nations history and that casts a negative light on both men who were very crucial to the founding of this country and our laws.

This paper was successful in solving the historical problem as both models predicted the same outcome, however there are additional classification models that can be used and might return a different result.