2020-0701 IST 707 Data Analytics

Homework #4

Sathish Kumar Rajendiran

2020-07-31

Table of Contents

[Task Summary 2](#_Toc47220951)

[Dataset Exploration 3](#_Toc47220952)

[Clustering 5](#_Toc47220953)

[Other Analysis 11](#_Toc47220954)

[Summary 13](#_Toc47220955)

[Conclusion 13](#_Toc47220956)

## **Task Summary**

This dataset consists of data from 85 federalist essays appeared in New York newspapers during 1787 - 1788 published under the pen name “Publius”. These are still considered as one of the most important sources for interpreting and understanding the original intent of the Constitution. There were 3 authors Hamilton, Madison and Jay had written most of these essays. Hamilton and Madison co-authored 3, Author Jay had written 5, Madison has 15 and Hamilton wrote 51 of them. However, remaining 11 are considered as disputed and claims were made to prove ownership of those essays, however none remain conclusive.

In this dataset, the features are a set of “function words” and it’s based on the percentage of word occurrence in an essay. Let’s try to solve this mystery by using clustering algorithms k-means, EM and HAC. In addition, provide evidences to support the claim as to who wrote the disputed essays, Hamilton or Madison? Task includes (but not limited to),

1. Dataset Exploration
   1. Data extraction
   2. Check for missing values
   3. Find NAs
2. Clustering
   1. K-means
   2. Euclidean mean
   3. HAC
   4. Dendrogram
3. Other Analytics
   1. Word cloud
4. Summary
   1. Conclude on who wrote the disputed essays, Hamilton or Madison.

## **Dataset Exploration**

In order to perform clustering analysis, first step is to understand the data and perform exploration analysis. It involves the following steps,

* 1. Data extraction - import necessary packages for data extraction.

A screenshot of a social media post

Description automatically generated

fedPapers85.csv – has 85 rows and 72 columns. 85 rows represent all the essays by their respective author categories.

A screenshot of a cell phone

Description automatically generated A screen shot of a computer

Description automatically generated

Preview and datatype of the dataset below,

A picture containing window, large, standing, flying

Description automatically generated

A close up of text on a white background

Description automatically generated

* 1. Check for missing values

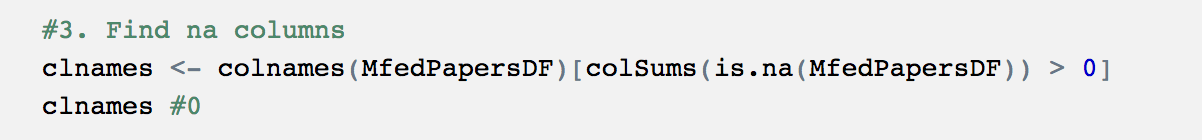
As part of the cleanup, first step is to look for any missing values in the dataset.

A picture containing knife

Description automatically generated

* 1. Find NAs

Next step is to find and replace any NAs from the dataset.



## **Clustering**

Next step is to perform clustering analysis by following the steps,

* **K-means**
* K-means with 3 clusters

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated A screenshot of a cell phone

Description automatically generated

From above, K-means with 3 clusters (possibly grouping 3 distinct authors work),

* function words from Jay’s essays are isolated in cluster number 3
* function words from “disputed” are both in cluster 1 and 2; similar to Hamilton’s, Madison’s.
* Hamilton & Madison’s co-authored essays are in cluster 2.
* Ratio of “between sum of squares” and “total sum of squares” is 27.1%
* K-means with 4 clusters

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated A screenshot of a cell phone

Description automatically generated

From above, K-means with 4 clusters - possibly 4 distinct groups (3 authors + combined work),

* function words from Jay’s essays are isolated in cluster number 4
* function words from “disputed” are both in cluster 2 and 3; similar to Hamilton’s, Madison’s and HM’s
* Ratio of “between sum of squares” and “total sum of squares” is 35.4%
* K-means (with 2 clusters)

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A picture containing text, map

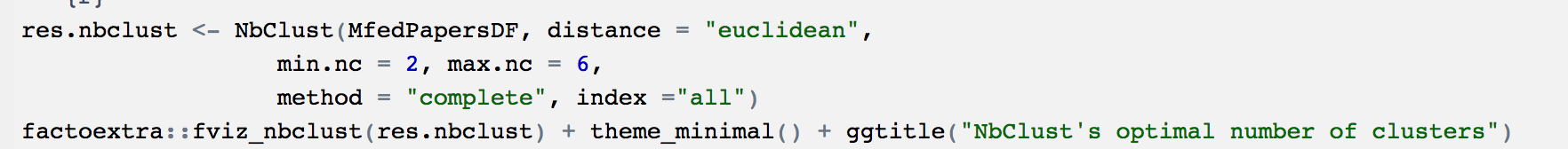
Description automatically generatedA screenshot of a cell phone

Description automatically generated

From above, K-means with 2 clusters - possibly to see major groupings,

* function words from Jay, Madison and Hamilton and half of disputed essays are isolated in cluster number 1
* majority of function words from disputed, Hamilton and Madison are in cluster 2.
* Ratio of “between sum of squares” and “total sum of squares” is 18.1%
* **Euclidean mean**

Let’s solve by running Euclidean mean and try to solve optimal number of clusters.



A screenshot of a cell phone

Description automatically generated

A close up of text on a white background

Description automatically generated

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

From above, it is obvious that 2, 5 and 4 are considered optimum cluster numbers for the clustering algorithm.

* **Elbow Chart**

Let’s look at another method to determine optimal number of clusters. Elbow method – is another well-known method in which the sum of squares at each number of clusters is calculated and change of slope from steep to shallow (elbow) is used to determine the optimal number of clusters.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generatedA screenshot of a cell phone

Description automatically generated

From above elbow chart, its confirmed that 2 is the optimal number of clusters.

* **Hierarchical Cluster Analysis (HAC)** – is alternative method to identifying groups in the dataset. It doesn’t require to pre-specify the number of clusters. In addition, it results in tree-based representation of observations called dendrogram.

**HAC with Complete Linkage**

A picture containing bird

Description automatically generated

A picture containing bird

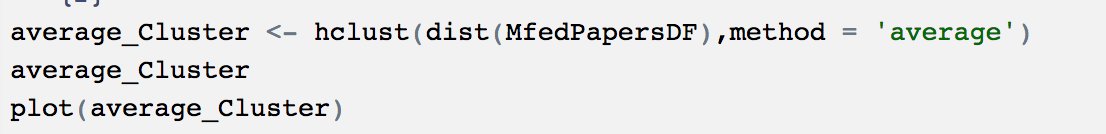
Description automatically generated

A close up of text on a white background

Description automatically generated

From the dendrogram above – Jay’s essays are grouped separately (Red box) and disputed essays are mixed with more of Madison and Hamilton’s (Green box). Also, mixed between co-authored essays. However, strong linkage is on Madison’s.

**HAC with Average Linkage**



A picture containing bird

Description automatically generated

A close up of text on a white background

Description automatically generated

From the dendrogram above – Jay’s essays are grouped separately (Red box) and disputed essays are mixed with more of Madison’s (Green box). Showing, strong linkage on Madison’s and clearly separating out Jay from the discussion.

## **Other Analysis**

Let’s perform Wordcloud on this dataset split by author’s work to identify most frequent words/pattern from these function words. Following words are removed from the analysis to increase the accuracy.

he|to|and|of|which|in.|by|be|that|it|a|is|as|will|would|for|from|or

A screenshot of a cell phone

Description automatically generated

A picture containing holding, person

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A picture containing drawing

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A picture containing holding, table, phone, person

Description automatically generated

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

A picture containing people

Description automatically generated

from Wordcloud analysis above,

* Hamilton can be excluded from the list
* It’s possible that Jay could have still authored it.
* Wordcloud also confirms strong correlation with Madison or Co-authoring with Madison doing most of writing.

## **Summary**

* Jay – possibly did not write those essays | **strong NO**
  + - K-means (with 2,3 and 4 clusters) ruling out the possibilities of Jay being the author. As Jay’s essays were grouped into different clusters most of the cases.
    - Hierarchical Clustering analysis (complete and average) confirm the same moving Jay’s into different group.
    - Wordcloud – didn’t help much in this case, though.
* Hamilton – possibly did not write those essays | **medium No**
  + - K-means (with 2,3 and 4 clusters) did not rule out the possibilities of Hamilton being the author. However, favors Hamilton & Madison co-authoring, though.
    - Hierarchical Clustering analysis (complete and average) confirms the same suggesting that Hamilton & Madison co-authoring and favors Madison’s more.
    - Wordcloud – clearly rule out Hamilton’s possibilities.

* HM (Co-authors) – likely possibility | **average Yes**
  + - K-means (with 2,3 and 4 clusters) did not rule out the possibilities of Hamilton being the author. However, favors Hamilton & Madison co-authoring, though.
    - Hierarchical Clustering analysis (complete and average) confirms the same suggesting that Hamilton & Madison co-authoring and favors Madison’s more.
    - Wordcloud – suggesting strong correlation on Hamilton & Madison co-authoring the essays.
* Madison – most likely possibility | **above average Yes**
  + - K-means (with 2,3 and 4 clusters) did not rule out the possibilities of Madison in all scenarios. There has been huge influence on Madison’s contribution suggesting likelihood of co-authoring.
    - Hierarchical Clustering analysis (complete and average) confirms the same suggesting that Hamilton & Madison co-authoring and favors Madison’s more.
    - Wordcloud – positive correlation on the possibilities of Hamilton & Madison co-authoring the essays. Also, suggesting Madison’s function words having more frequencies than others.

## **Conclusion**

Based on the clustering analysis involving K-means, Euclidean and Hierarchical clustering methods – None of these methods could not conclude any one author over other. However, strongly ruling out Jay being author of those disputed essays. Whereas favoring “Madison” as author having most influence with the words used. It is also possible that Hamilton & Madison could have co-authored adapting to Madison’s style of writing.

In addition, instead of just analyzing function words – may be sentence or phrases used in that essay analysis could help solve the mystery better.