IST 707- HW4

Alireza Zarrinmehr

2022-09-25

Using Clustering to Solve a Mystery in History

The Federalist Papers :

> The Federalist Papers were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay, the essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name "Publius." A bound edition of the essays was first published in 1788, but it was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

Disputed authorship:

> There are 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”. These are the famous essays with disputed authorship. Hamilton wrote to claim the authorship before he was killed in a duel. Later Madison also claimed authorship. Historians were trying to find out which one was the real author.

Collected data:

> We are provided with the Federalist Paper data set. The features are a set of “function words”, for example, “upon”. The feature value is the percentage of the word occurrence in an essay. For example, for the essay “Hamilton_fed_31.txt”, if the function word “upon” appeared 3 times, and the total number of words in this essay is 1000, the feature value is 3/1000=0.3%

Objectives:

> We are going to try solving this mystery and find the author of these articles.

Methods and limitations:

> Clustering algorithms k-Means and HAC will be implemented.

In K-Mean we to specify k, the number of clusters, in advance. This will not be a challenge in this problem as we know the number of authors.

HAC is sensitive to noise and outliers. However, we do not have an outlier in this data set

Analysis process:

Starting by loading the packages

```
require(stats)

## Loading required package: stats

require(cluster)

## Loading required package: cluster

require(dplyr)

## Loading required package: dplyr

require(ggvis)

## Loading required package: ggvis

require(tidyr)

## Loading required package: tidyr
```

Then continue by loading and the data file

```
#Loading the data file:
FedPapers <- read.csv('/Users/zarrinmehr/Desktop/11- IST.707.M002.FALL2
2.Applied Machine Learning 16797.1231/week 4/HW 4/HW4-data-fedPapers85.
csv')
#Getting the number of rows and columns:
dim(FedPapers)

## [1] 85 72

#The dataset comprises 85 observations and 72 variables.
```

Exploring the data file

We can see that the text's author and title are mentioned in the first two columns. The frequency of the letters that appear in the text is indicated in columns three and beyond.

```
#Looking at the first six rows:
head(FedPapers)
```

```
##   author         filename     a   all  also    an   and   any   are
as    at
## 1  dispt dispt_fed_49.txt 0.280 0.052 0.009 0.096 0.358 0.026 0.131
0.122 0.017
## 2  dispt dispt_fed_50.txt 0.177 0.063 0.013 0.038 0.393 0.063 0.051
0.139 0.114
## 3  dispt dispt_fed_51.txt 0.339 0.090 0.008 0.030 0.301 0.008 0.068
0.203 0.023
## 4  dispt dispt_fed_52.txt 0.270 0.024 0.016 0.024 0.262 0.056 0.064
0.111 0.056
## 5  dispt dispt_fed_53.txt 0.303 0.054 0.027 0.034 0.404 0.040 0.128
0.148 0.013
## 6  dispt dispt_fed_54.txt 0.245 0.059 0.007 0.067 0.282 0.052 0.111
0.252 0.015
##      be  been   but    by   can    do  down  even every  for.  from
had   has
## 1 0.411 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044
0.035 0.017
## 2 0.393 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101
0.101 0.013
## 3 0.474 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053
0.008 0.015
## 4 0.365 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079
0.016 0.024
## 5 0.344 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074
0.000 0.054
## 6 0.297 0.030 0.037 0.186 0.000 0.000 0.007 0.007 0.007 0.067 0.096
0.022 0.015
##    have her   his   if.   in.  into    is    it   its   may  more  m
ust my
## 1 0.044   0 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.
026  0
## 2 0.152   0 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.
013  0
## 3 0.023   0 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.
083  0
## 4 0.143   0 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.
071  0
## 5 0.047   0 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.
013  0
```

```
## 6 0.119   0 0.067 0.030 0.401 0.037 0.260 0.156 0.015 0.074 0.045 0.
015  0
##      no   not now    of    on   one  only    or   our shall should
so  some
## 1 0.035 0.114   0 0.900 0.140 0.026 0.035 0.096 0.017 0.017  0.017 0
.035 0.009
## 2 0.000 0.127   0 0.747 0.139 0.025 0.000 0.114 0.000 0.000  0.013 0
.013 0.063
## 3 0.030 0.068   0 0.858 0.150 0.030 0.023 0.060 0.000 0.008  0.068 0
.038 0.030
## 4 0.032 0.087   0 0.802 0.143 0.032 0.048 0.064 0.016 0.016  0.032 0
.040 0.024
## 5 0.047 0.128   0 0.869 0.054 0.047 0.027 0.081 0.027 0.000  0.000 0
.027 0.067
## 6 0.059 0.134   0 0.876 0.141 0.052 0.022 0.074 0.030 0.015  0.030 0
.007 0.045
##    such  than  that   the their  then there things  this    to up  u
pon   was
## 1 0.026 0.009 0.184 1.425 0.114 0.000 0.009  0.009 0.044 0.507  0 0.
000 0.009
## 2 0.000 0.000 0.152 1.254 0.165 0.000 0.000  0.000 0.051 0.355  0 0.
013 0.051
## 3 0.045 0.023 0.188 1.490 0.053 0.015 0.015  0.000 0.075 0.361  0 0.
000 0.008
## 4 0.008 0.000 0.238 1.326 0.071 0.008 0.000  0.000 0.103 0.532  0 0.
000 0.087
## 5 0.027 0.047 0.162 1.193 0.027 0.007 0.007  0.000 0.094 0.485  0 0.
000 0.027
## 6 0.015 0.030 0.208 1.469 0.089 0.007 0.007  0.000 0.126 0.445  0 0.
000 0.007
##    were  what  when which   who  will  with would your
## 1 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192    0
## 2 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139    0
## 3 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068    0
## 4 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064    0
## 5 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040    0
## 6 0.030 0.015 0.037 0.186 0.045 0.111 0.089 0.037    0
```

#We can see that the text's author and title are mentioned in the first two columns. The frequency of the letters that appear in the text is indicated in columns three and beyond.

#Getting the structure of data

```
str(FedPapers[,1:6])
```

```
## 'data.frame':    85 obs. of  6 variables:
##  $ author  : chr  "dispt" "dispt" "dispt" "dispt" ...
##  $ filename: chr  "dispt_fed_49.txt" "dispt_fed_50.txt" "dispt_fed_5
1.txt" "dispt_fed_52.txt" ...
##  $ a       : num  0.28 0.177 0.339 0.27 0.303 0.245 0.349 0.414 0.24
```

```
8 0.442 ...
##  $ all     : num  0.052 0.063 0.09 0.024 0.054 0.059 0.036 0.083 0.0
4 0.062 ...
##  $ also    : num  0.009 0.013 0.008 0.016 0.027 0.007 0.007 0.009 0.
007 0.006 ...
##  $ an      : num  0.096 0.038 0.03 0.024 0.034 0.067 0.029 0.018 0.0
4 0.075 ...
```

```
#Getting the types of variables:
table(sapply(FedPapers, class))
```

```
##
## character   numeric
##         2        70
```

```
#First two column are chracters and the rest are numeric.
```

```
#Check how many unique authors are there in the database:
unique(FedPapers$author)
```

```
## [1] "dispt"    "Hamilton" "HM"       "Jay"      "Madison"
```

```
#We'll choose three centroids to use as our starting point for the k-me
ans method. We selected three centroids since Hamilton, Madison, and Ja
y are each distinct writers.
```

##Data preparation:

```
#Creating a new data set by removing the first two columns of the origi
nal dataset:
FedPapersClean <- FedPapers[, 3:72]
```

```
#check the new data set
head(FedPapersClean)
```

```
##       a   all  also    an   and   any   are    as    at    be  been
but    by
## 1 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411 0.026
0.009 0.140
## 2 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393 0.165
0.000 0.139
## 3 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474 0.015
0.038 0.173
## 4 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365 0.127
0.032 0.167
## 5 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344 0.047
0.061 0.209
## 6 0.245 0.059 0.007 0.067 0.282 0.052 0.111 0.252 0.015 0.297 0.030
0.037 0.186
##     can    do  down  even every  for.  from   had   has  have her
his   if.
```

```
## 1 0.035 0.026 0.000 0.009 0.044 0.096 0.044 0.035 0.017 0.044   0 0.
017 0.000
## 2 0.000 0.013 0.000 0.025 0.000 0.076 0.101 0.101 0.013 0.152   0 0.
000 0.025
## 3 0.023 0.000 0.008 0.015 0.023 0.098 0.053 0.008 0.015 0.023   0 0.
000 0.023
## 4 0.056 0.000 0.000 0.024 0.040 0.103 0.079 0.016 0.024 0.143   0 0.
024 0.040
## 5 0.088 0.000 0.000 0.020 0.027 0.141 0.074 0.000 0.054 0.047   0 0.
020 0.034
## 6 0.000 0.000 0.007 0.007 0.007 0.067 0.096 0.022 0.015 0.119   0 0.
067 0.030
##     in.  into    is    it   its   may  more  must my    no   not now
of
## 1 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.026  0 0.035 0.114   0
0.900
## 2 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.013  0 0.000 0.127   0
0.747
## 3 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.083  0 0.030 0.068   0
0.858
## 4 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.071  0 0.032 0.087   0
0.802
## 5 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.013  0 0.047 0.128   0
0.869
## 6 0.401 0.037 0.260 0.156 0.015 0.074 0.045 0.015  0 0.059 0.134   0
0.876
##      on   one  only    or   our shall should    so  some  such  than
that
## 1 0.140 0.026 0.035 0.096 0.017 0.017  0.017 0.035 0.009 0.026 0.009
0.184
## 2 0.139 0.025 0.000 0.114 0.000 0.000  0.013 0.013 0.063 0.000 0.000
0.152
## 3 0.150 0.030 0.023 0.060 0.000 0.008  0.068 0.038 0.030 0.045 0.023
0.188
## 4 0.143 0.032 0.048 0.064 0.016 0.016  0.032 0.040 0.024 0.008 0.000
0.238
## 5 0.054 0.047 0.027 0.081 0.027 0.000  0.000 0.027 0.067 0.027 0.047
0.162
## 6 0.141 0.052 0.022 0.074 0.030 0.015  0.030 0.007 0.045 0.015 0.030
0.208
##     the their  then there things  this    to up  upon   was  were  w
hat  when
## 1 1.425 0.114 0.000 0.009  0.009 0.044 0.507  0 0.000 0.009 0.017 0.
000 0.009
## 2 1.254 0.165 0.000 0.000  0.000 0.051 0.355  0 0.013 0.051 0.000 0.
000 0.000
## 3 1.490 0.053 0.015 0.015  0.000 0.075 0.361  0 0.000 0.008 0.015 0.
008 0.000
## 4 1.326 0.071 0.008 0.000  0.000 0.103 0.532  0 0.000 0.087 0.079 0.
008 0.024
```

```
## 5 1.193 0.027 0.007 0.007  0.000 0.094 0.485  0 0.000 0.027 0.020 0.
020 0.007
## 6 1.469 0.089 0.007 0.007  0.000 0.126 0.445  0 0.000 0.007 0.030 0.
015 0.037
##   which   who  will  with would your
## 1 0.175 0.044 0.009 0.087 0.192    0
## 2 0.114 0.038 0.089 0.063 0.139    0
## 3 0.105 0.008 0.173 0.045 0.068    0
## 4 0.167 0.000 0.079 0.079 0.064    0
## 5 0.155 0.027 0.168 0.074 0.040    0
## 6 0.186 0.045 0.111 0.089 0.037    0
```

Cluster Analysis

After performing a preliminary analysis of the data, we can move on to evaluate various clustering techniques that could provide information about the author(s).

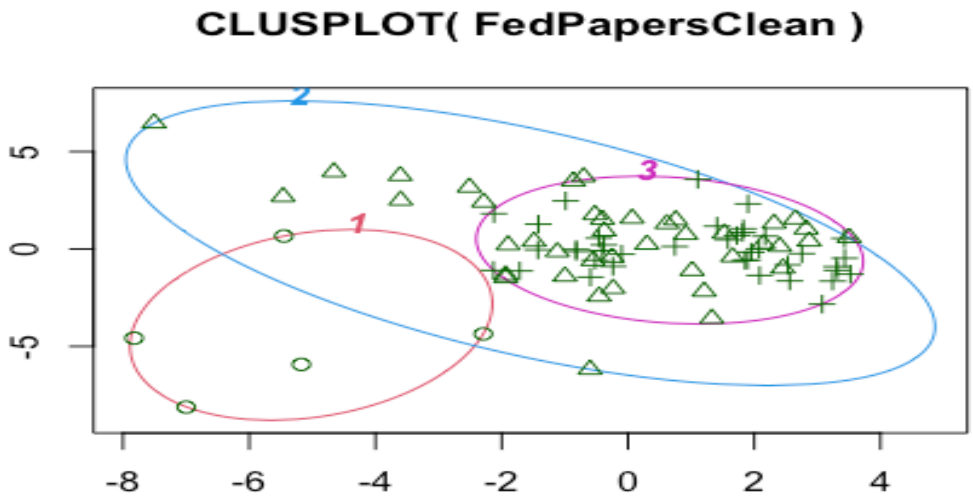
K-means and HAC are two clustering analyses that we will do.

1-1 K-Mean

We'll choose three centroids to use as our starting point for the k-means method. We selected three centroids since Hamilton, Madison, and Jay are each distinct writer.

```
#initiate clustering with k = 3
set.seed(54)
KMeans <- kmeans(FedPapersClean, 3)

#Ploting the clusters
clusplot(FedPapersClean, KMeans$cluster, color = T, shade = F, labels =
5, lines = 0
```

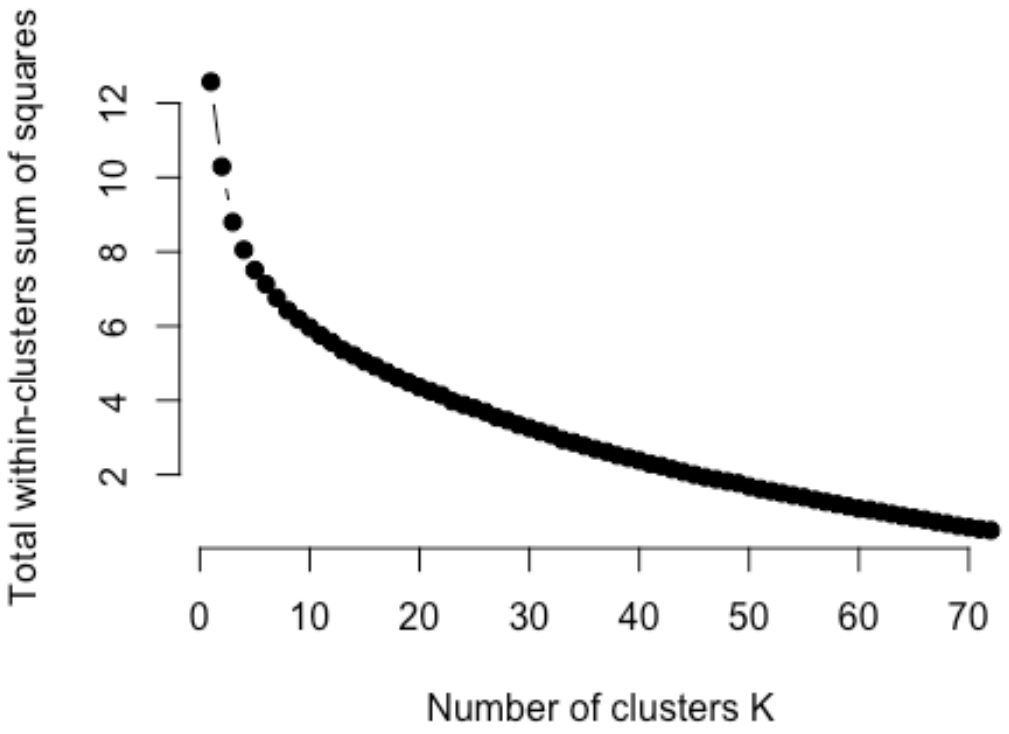


As can be seen in the figure, the four groups overlap one another. A *better strategy is making* an elbow chart to establish the appropriate number of centroids.

1-2 Elbow method

```
#Elbow Method for finding the optimal number of clusters
set.seed(100)
# Compute and plot wss for k = 2 to k = 15.

wss <- sapply(1:72, function(k){kmeans(FedPapersClean, k, nstart=50,ite
r.max = 72 )$tot.withinss})
plot(1:72, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```



There is not a clear breakpoint (elbow) in the chart. Thus, we continue with K of 3 which (in this case) makes the most sense.

```
FedsCluster <- cbind(FedPapers, KMeans$cluster)
colnames(FedsCluster)[73] <- 'cluster'
FedsClusterClean <- FedsCluster %>% group_by(author, cluster) %>% summa
rise(number = n())

## `summarise()` has grouped output by 'author'. You can override using
the
## `.groups` argument.

FedsClusterClean

## # A tibble: 8 × 3
## # Groups:   author [5]
##   author   cluster number
##   <chr>      <int>  <int>
## 1 dispt          2      6
## 2 dispt          3      5
## 3 Hamilton       2     25
## 4 Hamilton       3     26
## 5 HM             2      3
## 6 Jay            1      5
## 7 Madison        2      7
## 8 Madison        3      8

spread(FedsClusterClean, key = cluster, value = number)

## # A tibble: 5 × 4
## # Groups:   author [5]
##   author     `1`   `2`   `3`
##   <chr>    <int> <int> <int>
## 1 dispt       NA     6     5
## 2 Hamilton    NA    25    26
## 3 HM          NA     3    NA
## 4 Jay          5    NA    NA
## 5 Madison     NA     7     8
```

Disputed texts are clustered in clusters 2 and 3. Almost half of Hamilton's writings are clustered in cluster 2, and half of it is clustered in cluster 3. 7 of Madison's writings are clustered in cluster 2, and 8 of his texts are clustered in cluster 3. In cluster 1 we have all of Jay's writing.

Therefore, we do not have enough evidence to support or deny any claims on the disputed texts.

2-1 HAC (complete method): Looking at clustering through a hierarchical clustering algorithm

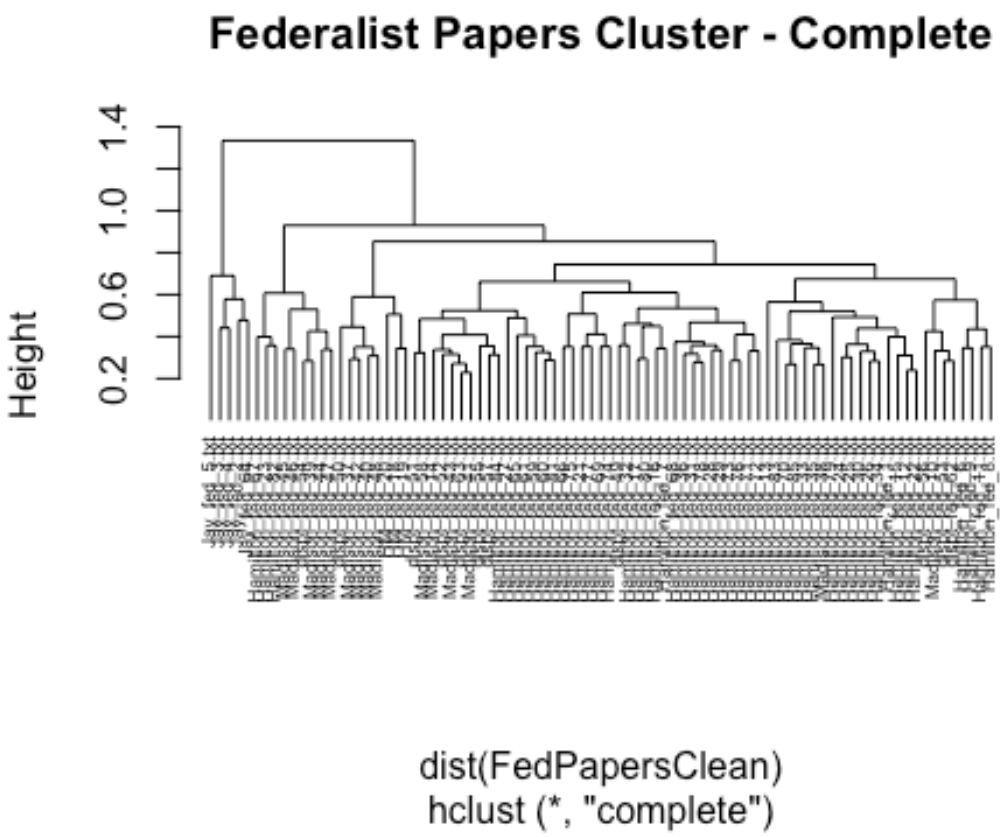
```
#Building two clusters using the complete method and the averag method
HACComp <- hclust(dist(FedPapersClean), method = 'complete')
HACAvg <- hclust(dist(FedPapersClean), method = 'average')
#Calling both HAC clustering
HACComp

##
## Call:
## hclust(d = dist(FedPapersClean), method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 85

HACAvg

##
## Call:
## hclust(d = dist(FedPapersClean), method = "average")
##
## Cluster method   : average
## Distance         : euclidean
## Number of objects: 85

#Graph the clusters to compare HAC and k-means algorithm
plot(HACComp, hang = -1, cex = 0.5, main = "Federalist Papers Cluster -
Complete", label = FedPapers$filename)
```

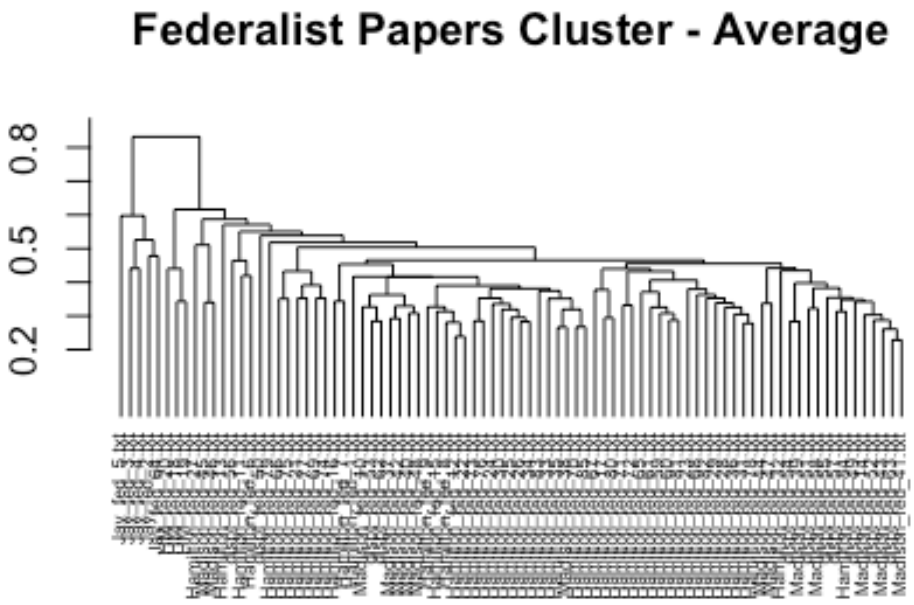


Looking at the chart, the following can be interpreted:

- All of Jay's papers are arranged in a single group to the left of the dendrogram.
- The co-authored (HM) papers are toward the left side of the dendrogram, clustered with Madison's papers.
- Some of the disputed files (54, 50, 51, 52, 63) are grouped with Madison's files and may be attributed to Madison.
- Some of the disputed files can not be strongly attributed to just one author (55, 57, 56, 53, 62).
- Some of the disputed files (49) are grouped with Hamilton's files and may be attributed to Hamilton.

2-2 HAC (average method): Looking at clustering through a hierarchical clustering algorithm

```
plot(HACAvg, hang = -1, cex = 0.5, main = "Federalist Papers Cluster -
Average", label = FedPapers$filename)
```



Looking at the chart, the following can be interpreted:

- All of Jay's papers are arranged in a single group to the left of the dendrogram.
- The co-authored (HM) papers are at the left side of the dendrogram, clustered with Madison's papers.
- Some of the disputed files (54, 51, 55, 57, 52, 63, 53, 62) are grouped with Madison's files and may be attributed to Madison.
- Some of the disputed files can not be strongly attributed to just one author (49).
- Some of the disputed files (50, 56) are grouped with Hamilton's files and may be attributed to Hamilton.

Conclusions

There in no consensus on the result of clustering models. The Hamilton, Madison, and contested publications were ultimately clustered so closely together that we were unable to determine any real authorship. However, there may be a chance that texts 51, 52, 54, and 63 could be attributed to Madison, and 56 could be attributed to Hamilton as both of the HAC clustering models (average method and complete method) agree on that.

Possibly, these 11 papers are a combination of co-authored works or a very realistic attempt by both authors to replicate each other's writing style, which makes it even harder to determine who the genuine author is.