

Determining Authorship of the Disputed Federalist Papers using K-Means and HAC Clustering Algorithms

Sarah Morris
2024-02-08

Overview

1. About the Federalist Papers Quote from the Library of Congress
<http://www.loc.gov/rr/program/bib/ourdocs/federalist.html> 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.
2. About the disputed authorship The original essays can be downloaded from the Library of Congress. <http://thomas.loc.gov/home/histdox/fedpapers.html> In the author column, you 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”. 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.
3. Computational approach for authorship attribution In 1960s, statistician Mosteller and Wallace analyzed the frequency distributions of common function words in the Federalist Papers, and drew their conclusions. This is a pioneering work on using mathematical approaches for authorship attribution.
<http://www.stat.cmu.edu/~vlachos/courses/724/final/mosteller.pdf> Nowadays, authorship attribution has become a classic problem in the data mining field, with applications in forensics (e.g. deception detection), and information organization.

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 $3/1000=0.3\%$

Now I will try solving this mystery using clustering algorithms k-Means and HAC.

Read in Dataset

```
fedpapers_orig<-read.csv("~/Downloads/week4_resources/fedPapers85_fromClass.csv",
na.strings = c(""))

# check for missing values
sum(is.na(fedpapers_orig))

## [1] 0
```

Read in all Necessary Libraries

```
library(tidyverse)

## — Attaching core tidyverse packages ————— tidyverse 2.0.0 —
## ✓ dplyr   1.1.3   ✓ readr   2.1.4
## ✓ forcats 1.0.0   ✓ stringr 1.5.1
## ✓ ggplot2 3.4.4   ✓ tibble  3.2.1
## ✓ lubridate 1.9.3 ✓ tidyr   1.3.0
## ✓ purrr   1.0.2

## — Conflicts ————— tidyverse_conflicts()
—
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ⓘ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
#install.packages("factoextra")
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(stats)
```

Exploratory Data Analysis

Firstly, let's remove the papers written by John Jay, as we already know that the disputed papers must be written by either Madison or Hamilton:

```
fedpapers<-fedpapers_orig[fedpapers_orig$author != "Jay",]
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 must 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 upon 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
```

Let's group by author and try to understand the variance in writing styles between them.

```
#deleting file name
fedpapers_eda<-fedpapers[,-2]
#Take mean frequency of each group grouped by author

group_fedpapers<- fedpapers_eda %>% group_by(author) %>%
  summarise(across(everything(), mean),
    .groups = 'drop') %>%
  as.data.frame()
group_fedpapers
```

author a all also an and any
1 HM 0.2133333 0.04266667 0.006000000 0.04700000 0.5306667 0.01800000
2 Hamilton 0.3156078 0.05376471 0.004784314 0.08080392 0.3394902 0.04674510
3 Madison 0.2698000 0.05533333 0.011066667 0.05946667 0.4196667 0.02980000
4 dispt 0.3039091 0.05554545 0.011272727 0.04845455 0.3560000 0.04218182
are as at be been but by
1 0.08233333 0.0800000 0.04433333 0.06666667 0.02133333 0.02833333 0.1750000
2 0.07254902 0.1177255 0.04882353 0.30837255 0.06192157 0.03015686 0.1045882
3 0.07486667 0.1340667 0.02953333 0.28953333 0.07406667 0.03540000 0.1660667
4 0.09590909 0.1380000 0.04700000 0.35918182 0.05500000 0.03154545 0.1615455
can do down even every for.
1 0.005333333 0.002333333 0.000000000 0.002333333 0.007666667 0.08566667
2 0.038235294 0.006607843 0.0019803922 0.012294118 0.020803922 0.09245098
3 0.028333333 0.004000000 0.0009333333 0.009800000 0.031066667 0.09453333
4 0.042545455 0.007909091 0.0013636364 0.013636364 0.041090909 0.10000000
from had has have her his
1 0.10866667 0.08133333 0.04600000 0.06666667 0.020333333 0.08700000
2 0.08021569 0.01743137 0.04811765 0.09894118 0.007058824 0.03192157
3 0.06853333 0.02246667 0.05106667 0.10046667 0.007800000 0.01886667
4 0.08009091 0.02245455 0.02490909 0.07872727 0.006909091 0.01963636
if. in. into is it its may
1 0.007666667 0.249000 0.02966667 0.1040000 0.1086667 0.05400000 0.01733333
2 0.029019608 0.343902 0.02131373 0.1594118 0.1585882 0.05254902 0.06219608
3 0.021466667 0.286400 0.02686667 0.1714000 0.1498000 0.04806667 0.06560000
4 0.021363636 0.276000 0.02236364 0.1638182 0.1489091 0.03463636 0.06927273
more must my no not now
1 0.04200000 0.01366667 0.0053333333 0.02400000 0.03466667 0.012666667
2 0.03996078 0.03370588 0.0042549020 0.03188235 0.09029412 0.006549020
3 0.04953333 0.03380000 0.0018666667 0.04280000 0.09500000 0.005133333
4 0.04872727 0.03963636 0.0006363636 0.03054545 0.10790909 0.002818182
of on one only or our shall
1 0.8386667 0.09366667 0.04400000 0.01500000 0.05366667 0.00600000 0.004666667
2 0.9571765 0.04743137 0.03568627 0.02031373 0.09801961 0.02266667 0.021705882

```

## 3 0.8687333 0.10506667 0.04280000 0.02360000 0.08420000 0.01786667 0.01986667
## 4 0.8851818 0.11254545 0.04236364 0.02663636 0.09054545 0.01663636 0.008000000
##   should      so      some      such      than      that      the
## 1 0.002333333 0.01733333 0.02133333 0.03100000 0.03766667 0.1016667 1.267000
## 2 0.029392157 0.02896078 0.01601961 0.02894118 0.04454902 0.2211765 1.289941
## 3 0.023733333 0.02900000 0.02133333 0.02873333 0.04100000 0.2015333 1.375267
## 4 0.017181818 0.03163636 0.03481818 0.02072727 0.03845455 0.1998182 1.307000
##   their      then      there      things      this      to
## 1 0.13233333 0.009666667 0.004333333 0.000000000 0.06500000 0.3826667
## 2 0.07468627 0.005196078 0.037098039 0.003372549 0.09349020 0.5910784
## 3 0.08913333 0.006200000 0.007733333 0.001400000 0.08313333 0.4568667
## 4 0.09263636 0.008181818 0.013727273 0.002363636 0.08363636 0.4526364
##   up      upon      was      were      what      when
## 1 0.009000000 0.005333333 0.11400000 0.04733333 0.002333333 0.01566667
## 2 0.004568627 0.047313725 0.02060784 0.01745098 0.013627451 0.01225490
## 3 0.001266667 0.002200000 0.02573333 0.02086667 0.011600000 0.00600000
## 4 0.001545455 0.001181818 0.02663636 0.02090909 0.011363636 0.01190909
##   which      who      will      with      would      your
## 1 0.1373333 0.04733333 0.01666667 0.09700000 0.02566667 0.000000000
## 2 0.1604902 0.03292157 0.09239216 0.07896078 0.12272549 0.002078431
## 3 0.1638000 0.02653333 0.10666667 0.07706667 0.06066667 0.002266667
## 4 0.1692727 0.02618182 0.12663636 0.07490909 0.07036364 0.000000000

```

Choose “meaningful” words to analyze

As we know, some words are more meaningful than others. Let’s see who writes most frequently in a first and second person tone (an informal tone, call to action!). Therefore, let’s look at the relative frequencies of “my”, “our”, and “your” based on the author.

```

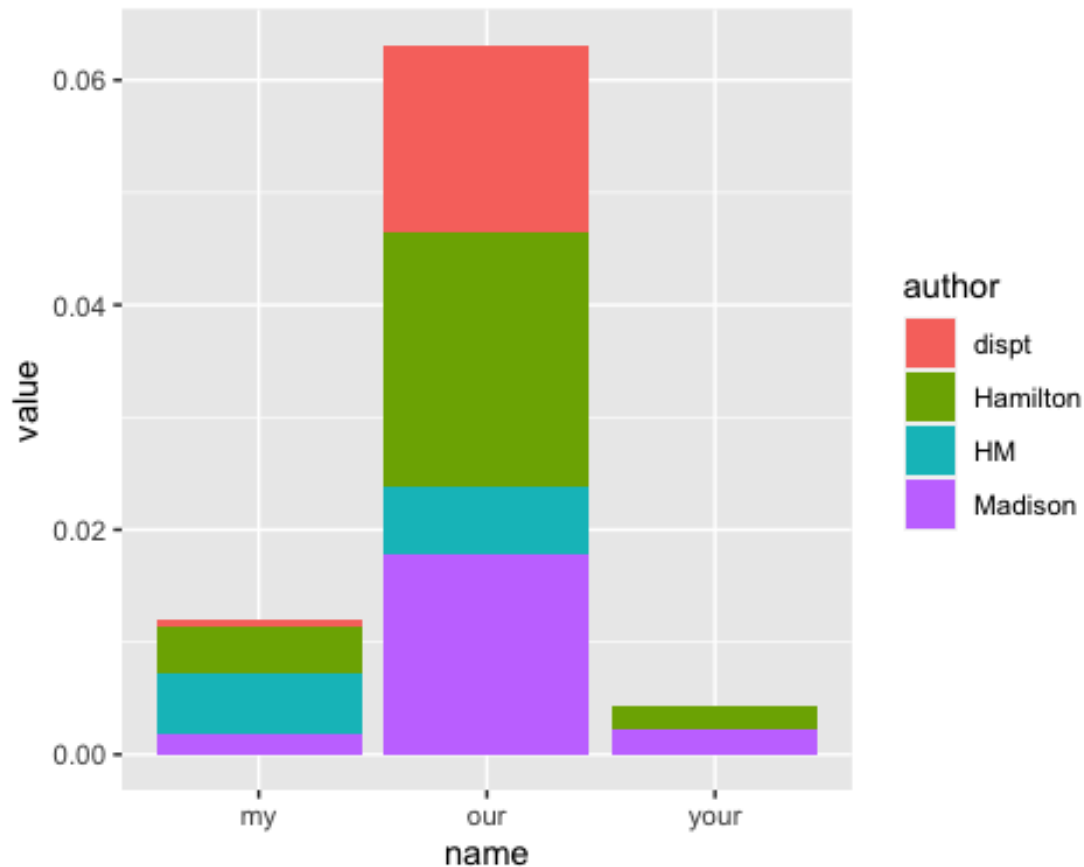
informal<-select(group_fedpapers, c('author','my', 'our', 'your'))
informal_chart<-pivot_longer(informal, -author)
informal_chart

## # A tibble: 12 × 3
##   author name    value

```

```
##  <chr>  <chr>  <dbl>
##  1 HM    my    0.00533
##  2 HM    our   0.006
##  3 HM    your  0
##  4 Hamilton my  0.00425
##  5 Hamilton our 0.0227
##  6 Hamilton your 0.00208
##  7 Madison my  0.00187
##  8 Madison our 0.0179
##  9 Madison your 0.00227
## 10 dispt  my  0.000636
## 11 dispt  our 0.0166
## 12 dispt  your 0
```

```
ggplot(data=informal_chart, aes(x = name, y=value, fill = author)) +  
  geom_col(position=position_stack())
```



Based on the above chart, Hamilton appears to write more frequently with the words “my” and “our”, which gives him a more personal tone and a more communal feel with his audience. Both Hamilton and Madison use the word “your” in relatively equal frequencies. Strangely, the disputed papers do not use the word “your” at all.

“Our” is used much more frequently than the words “my” and “your”, suggesting that both Madison and Hamilton used “us” and “we” verbiage as a way to relate to their audience as Americans and call them to action.

Let’s follow the idea of “call to action” words in our exploratory data analysis. It is intuitive that one writer may have a more powerful, or strong writing style. Certain words imply urgency, or action, such as the following: must, now, shall, and should. Let’s examine the relative frequency of each for each author:

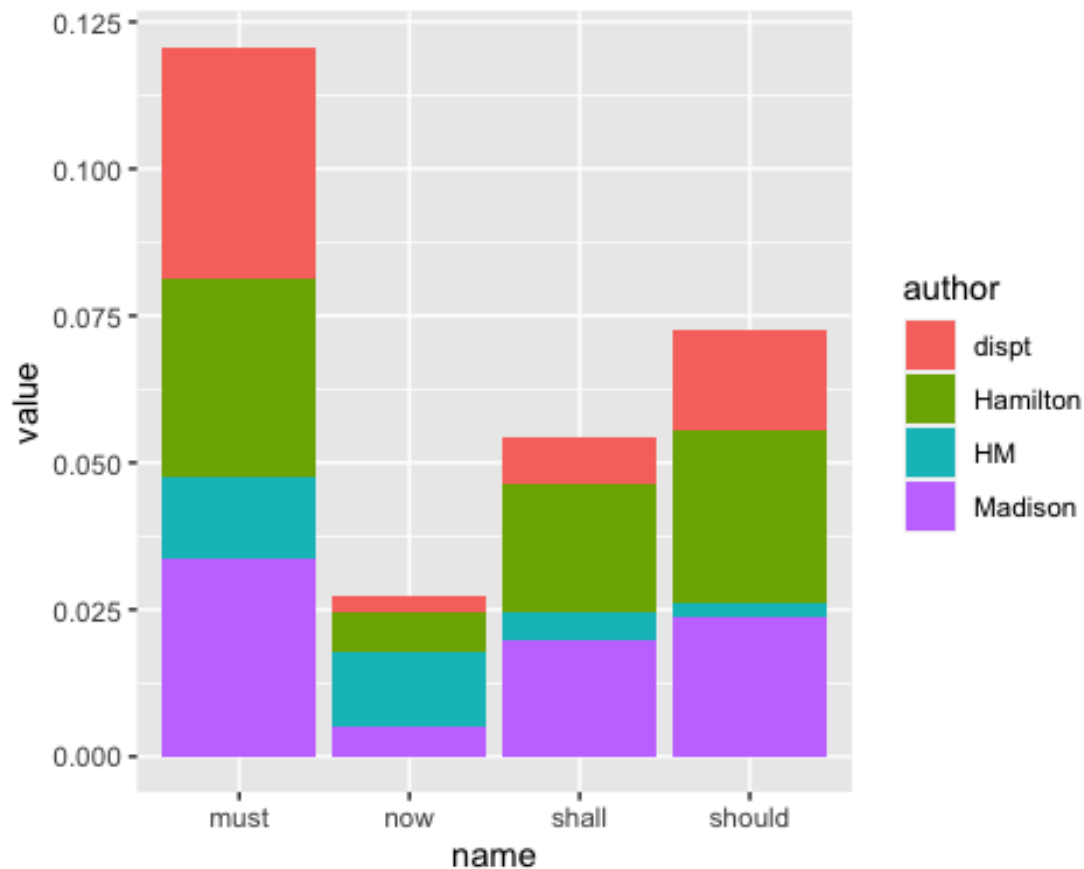
```
action<-select(group_fedpapers, c('author','must', 'now', 'shall', 'should'))
action_chart<-pivot_longer(action, -author)
action_chart

## # A tibble: 16 × 3
##   author name  value
##   <chr>  <chr>  <dbl>
```



```
## 1 HM    must 0.0137
## 2 HM    now  0.0127
## 3 HM    shall 0.00467
## 4 HM    should 0.00233
## 5 Hamilton must 0.0337
## 6 Hamilton now 0.00655
## 7 Hamilton shall 0.0217
## 8 Hamilton should 0.0294
## 9 Madison must 0.0338
## 10 Madison now 0.00513
## 11 Madison shall 0.0199
## 12 Madison should 0.0237
## 13 dispt must 0.0396
## 14 dispt now 0.00282
## 15 dispt shall 0.008
## 16 dispt should 0.0172
```

```
ggplot(data=action_chart, aes(x = name, y=value, fill = author)) +  
  geom_col(position=position_stack())
```



Again, Hamilton seems to use these “call to action” words slightly more frequently. This provides a bit of preliminary context to our analysis regarding the tones of each of our writers, which may be useful in identifying the author of the disputed papers.

K-Means Algorithm

Now we will use our K-means algorithm to sort the documents into clusters.

```
#remove author name
fedpapers_km<-fedpapers[,-1]
rownames(fedpapers_km)<-fedpapers_km[,1]
#view new dataframe with each row labeled by file name.
fedpapers_km[,1]<-NULL
```

Run K-Means Clustering

```
#install.packages("stats")
library(stats)
library(factoextra)
```

```

#set seed
set.seed(22)
#drop author names for cluster
#fedpapers_km2<-fedpapers[,-1]
cluster<-kmeans(fedpapers_km, 5, nstart = 25)
fedpapers_km$clusters<-as.factor(cluster$cluster)
fedpapers$clusters<-as.factor(cluster$cluster)
#view results
print(cluster)

## K-means clustering with 5 clusters of sizes 14, 13, 22, 16, 15
##
## Cluster means:
##      a      all      also      an      and      any      are
## 1 0.3465000 0.04557143 0.008142857 0.09221429 0.3001429 0.05385714 0.05207143
## 2 0.2590000 0.05815385 0.012076923 0.05046154 0.3574615 0.03469231 0.08646154
## 3 0.3323636 0.05227273 0.004181818 0.07077273 0.3445909 0.04422727 0.07872727
## 4 0.2750625 0.06356250 0.003187500 0.08168750 0.3610000 0.04700000 0.08212500
## 5 0.2796667 0.05000000 0.009200000 0.05840000 0.4607333 0.02793333 0.08173333
##      as      at      be      been      but      by      can
## 1 0.1197143 0.04242857 0.3321429 0.06492857 0.02914286 0.1065714 0.03400000
## 2 0.1513077 0.03253846 0.3343077 0.05592308 0.03169231 0.1576923 0.03200000
## 3 0.1081818 0.06186364 0.3152273 0.05745455 0.02945455 0.1191818 0.03422727
## 4 0.1242500 0.04000000 0.2983125 0.06062500 0.03200000 0.0886875 0.05131250
## 5 0.1174667 0.03766667 0.2344667 0.07120000 0.03473333 0.1696000 0.02620000
##      do      down      even      every      for.      from
## 1 0.006785714 0.001000000 0.00800000 0.01964286 0.09342857 0.07971429
## 2 0.004846154 0.0019230769 0.01076923 0.03638462 0.09115385 0.06476923
## 3 0.006909091 0.001500000 0.01504545 0.02840909 0.09736364 0.08245455
## 4 0.006562500 0.003375000 0.01231250 0.01818750 0.08731250 0.07918750
## 5 0.005066667 0.0002666667 0.01006667 0.02253333 0.09720000 0.08580000
##      had      has      have      her      his      if.      in.
## 1 0.01757143 0.04292857 0.09335714 0.000500000 0.02307143 0.03057143 0.3887143
## 2 0.01507692 0.03784615 0.09115385 0.004538462 0.01615385 0.02084615 0.2868462

```

3 0.01840909 0.05427273 0.09436364 0.01563636 0.03804545 0.02981818 0.3210000
4 0.01406250 0.04156250 0.09887500 0.001687500 0.03106250 0.02681250 0.3203750
5 0.04300000 0.04533333 0.09793333 0.011800000 0.03473333 0.01840000 0.2839333
into is it its may more must
1 0.01814286 0.1677143 0.1842143 0.04907143 0.06692857 0.03557143 0.02578571
2 0.02138462 0.1710000 0.1536154 0.04661538 0.06346154 0.05323077 0.03423077
3 0.02509091 0.1542727 0.1370909 0.04695455 0.05850000 0.04781818 0.03722727
4 0.01681250 0.1723125 0.1655625 0.06118750 0.06956250 0.03706250 0.04025000
5 0.03146667 0.1395333 0.1372000 0.04260000 0.05386667 0.04053333 0.02893333
my no not now of on one
1 0.003214286 0.03900000 0.09550000 0.010142857 1.0425000 0.04671429 0.03228571
2 0.002153846 0.04423077 0.09523077 0.004000000 0.8984615 0.12669231 0.04269231
3 0.004909091 0.02990909 0.09359091 0.005727273 0.9251818 0.05968182 0.04177273
4 0.004437500 0.03050000 0.08156250 0.004875000 0.9122500 0.03793750 0.03325000
5 0.001066667 0.02726667 0.09213333 0.005466667 0.8583333 0.08620000 0.04013333
only or our shall should so some
1 0.02364286 0.09650000 0.005428571 0.02564286 0.02742857 0.02885714 0.01757143
2 0.02623077 0.08953846 0.015000000 0.01838462 0.02353846 0.02507692 0.02192308
3 0.01622727 0.09686364 0.040363636 0.01986364 0.02618182 0.02845455 0.01754545
4 0.02250000 0.10112500 0.015000000 0.01943750 0.03400000 0.03293750 0.01181250
5 0.02260000 0.07700000 0.015066667 0.01073333 0.01606667 0.02860000 0.03186667
such than that the their then there
1 0.02850000 0.04628571 0.2338571 1.394857 0.05842857 0.004285714 0.037428571
2 0.02369231 0.03646154 0.1969231 1.481538 0.08430769 0.006000000 0.007769231
3 0.02854545 0.05200000 0.2208636 1.147364 0.08272727 0.005954545 0.034227273
4 0.03006250 0.03693750 0.2151875 1.390937 0.07937500 0.006312500 0.038500000
5 0.02746667 0.03773333 0.1780000 1.220600 0.10386667 0.007133333 0.011866667
things this to up upon was
1 0.003071429 0.10928571 0.5442857 0.0011428571 0.04771429 0.01642857
2 0.002000000 0.08276923 0.4475385 0.0004615385 0.000000000 0.02607692
3 0.003863636 0.09559091 0.5812273 0.0084090909 0.04372727 0.01727273
4 0.003500000 0.07737500 0.6515000 0.0024375000 0.04718750 0.02412500
5 0.000600000 0.07886667 0.4317333 0.0033333333 0.00600000 0.04913333

```

##      were      what      when      which      who      will      with
## 1 0.01700000 0.01650000 0.017357143 0.1444286 0.02971429 0.08950000 0.07550000
## 2 0.01946154 0.011230769 0.010230769 0.1620000 0.02123077 0.12584615 0.07546154
## 3 0.01754545 0.010000000 0.013818182 0.1611364 0.03054545 0.09254545 0.08077273
## 4 0.01625000 0.016875000 0.006062500 0.1800000 0.03718750 0.10200000 0.07937500
## 5 0.02920000 0.008933333 0.007733333 0.1575333 0.03653333 0.07986667 0.08086667
##      would      your
## 1 0.15028571 0.000000000
## 2 0.07838462 0.002615385
## 3 0.13818182 0.003818182
## 4 0.07356250 0.001375000
## 5 0.04533333 0.000000000
##
## Clustering vector:
##  dispt_fed_49.txt  dispt_fed_50.txt  dispt_fed_51.txt  dispt_fed_52.txt
##           2           5           2           2
##  dispt_fed_53.txt  dispt_fed_54.txt  dispt_fed_55.txt  dispt_fed_56.txt
##           5           2           3           5
##  dispt_fed_57.txt  dispt_fed_62.txt  dispt_fed_63.txt  Hamilton_fed_1.txt
##           2           3           5           3
## Hamilton_fed_11.txt Hamilton_fed_12.txt Hamilton_fed_13.txt Hamilton_fed_15.txt
##           3           3           3           3
## Hamilton_fed_16.txt Hamilton_fed_17.txt Hamilton_fed_21.txt Hamilton_fed_22.txt
##           1           4           1           3
## Hamilton_fed_23.txt Hamilton_fed_24.txt Hamilton_fed_25.txt Hamilton_fed_26.txt
##           4           3           3           3
## Hamilton_fed_27.txt Hamilton_fed_28.txt Hamilton_fed_29.txt Hamilton_fed_30.txt
##           4           4           4           3
## Hamilton_fed_31.txt Hamilton_fed_32.txt Hamilton_fed_33.txt Hamilton_fed_34.txt
##           4           1           4           3
## Hamilton_fed_35.txt Hamilton_fed_36.txt Hamilton_fed_59.txt  Hamilton_fed_6.txt
##           3           4           1           5
## Hamilton_fed_60.txt Hamilton_fed_61.txt Hamilton_fed_65.txt Hamilton_fed_66.txt

```

```

##           1           1           1           1
## Hamilton_fed_67.txt Hamilton_fed_68.txt Hamilton_fed_69.txt Hamilton_fed_7.txt
##           4           4           1           3
## Hamilton_fed_70.txt Hamilton_fed_71.txt Hamilton_fed_72.txt Hamilton_fed_73.txt
##           4           4           3           3
## Hamilton_fed_74.txt Hamilton_fed_75.txt Hamilton_fed_76.txt Hamilton_fed_77.txt
##           1           1           3           1
## Hamilton_fed_78.txt Hamilton_fed_79.txt Hamilton_fed_8.txt Hamilton_fed_80.txt
##           4           3           3           4
## Hamilton_fed_81.txt Hamilton_fed_82.txt Hamilton_fed_83.txt Hamilton_fed_84.txt
##           1           4           3           1
## Hamilton_fed_85.txt Hamilton_fed_9.txt HM_fed_18.txt HM_fed_19.txt
##           4           3           5           5
## HM_fed_20.txt Madison_fed_10.txt Madison_fed_14.txt Madison_fed_37.txt
##           5           5           2           5
## Madison_fed_38.txt Madison_fed_39.txt Madison_fed_40.txt Madison_fed_41.txt
##           5           2           5           5
## Madison_fed_42.txt Madison_fed_43.txt Madison_fed_44.txt Madison_fed_45.txt
##           5           2           2           2
## Madison_fed_46.txt Madison_fed_47.txt Madison_fed_48.txt Madison_fed_58.txt
##           2           2           5           2
##
## Within cluster sum of squares by cluster:
## [1] 1.182484 1.047539 1.782588 1.191858 1.325325
## (between_SS / total_SS = 32.8 %)
##
## Available components:
##
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"
## [6] "betweenss" "size" "iter" "ifault"

head(fedpapers_km)

```

```

##          a all also an and any are as at be
## dispt_fed_49.txt 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411
## dispt_fed_50.txt 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393
## dispt_fed_51.txt 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474
## dispt_fed_52.txt 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365
## dispt_fed_53.txt 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344
## dispt_fed_54.txt 0.245 0.059 0.007 0.067 0.282 0.052 0.111 0.252 0.015 0.297
##          been but by can do down even every for. from
## dispt_fed_49.txt 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044
## dispt_fed_50.txt 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101
## dispt_fed_51.txt 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053
## dispt_fed_52.txt 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079
## dispt_fed_53.txt 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074
## dispt_fed_54.txt 0.030 0.037 0.186 0.000 0.000 0.007 0.007 0.007 0.067 0.096
##          had has have her his if. in. into is it
## dispt_fed_49.txt 0.035 0.017 0.044 0 0.017 0.000 0.262 0.009 0.157 0.175
## dispt_fed_50.txt 0.101 0.013 0.152 0 0.000 0.025 0.291 0.025 0.038 0.127
## dispt_fed_51.txt 0.008 0.015 0.023 0 0.000 0.023 0.308 0.038 0.150 0.173
## dispt_fed_52.txt 0.016 0.024 0.143 0 0.024 0.040 0.238 0.008 0.151 0.222
## dispt_fed_53.txt 0.000 0.054 0.047 0 0.020 0.034 0.263 0.013 0.189 0.108
## dispt_fed_54.txt 0.022 0.015 0.119 0 0.067 0.030 0.401 0.037 0.260 0.156
##          its may more must my no not now of on one
## dispt_fed_49.txt 0.070 0.035 0.026 0.026 0 0.035 0.114 0 0.900 0.140 0.026
## dispt_fed_50.txt 0.038 0.038 0.000 0.013 0 0.000 0.127 0 0.747 0.139 0.025
## dispt_fed_51.txt 0.030 0.120 0.038 0.083 0 0.030 0.068 0 0.858 0.150 0.030
## dispt_fed_52.txt 0.048 0.056 0.056 0.071 0 0.032 0.087 0 0.802 0.143 0.032
## dispt_fed_53.txt 0.013 0.047 0.067 0.013 0 0.047 0.128 0 0.869 0.054 0.047
## dispt_fed_54.txt 0.015 0.074 0.045 0.015 0 0.059 0.134 0 0.876 0.141 0.052
##          only or our shall should so some such than that
## dispt_fed_49.txt 0.035 0.096 0.017 0.017 0.017 0.035 0.009 0.026 0.009 0.184
## dispt_fed_50.txt 0.000 0.114 0.000 0.000 0.013 0.013 0.063 0.000 0.000 0.152
## dispt_fed_51.txt 0.023 0.060 0.000 0.008 0.068 0.038 0.030 0.045 0.023 0.188
## dispt_fed_52.txt 0.048 0.064 0.016 0.016 0.032 0.040 0.024 0.008 0.000 0.238

```

```

## dispt_fed_53.txt 0.027 0.081 0.027 0.000 0.000 0.027 0.067 0.027 0.047 0.162
## dispt_fed_54.txt 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
## dispt_fed_49.txt 1.425 0.114 0.000 0.009 0.009 0.044 0.507 0 0.000 0.009
## dispt_fed_50.txt 1.254 0.165 0.000 0.000 0.000 0.051 0.355 0 0.013 0.051
## dispt_fed_51.txt 1.490 0.053 0.015 0.015 0.000 0.075 0.361 0 0.000 0.008
## dispt_fed_52.txt 1.326 0.071 0.008 0.000 0.000 0.103 0.532 0 0.000 0.087
## dispt_fed_53.txt 1.193 0.027 0.007 0.007 0.000 0.094 0.485 0 0.000 0.027
## dispt_fed_54.txt 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 clusters
## dispt_fed_49.txt 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192 0 2
## dispt_fed_50.txt 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139 0 5
## dispt_fed_51.txt 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068 0 2
## dispt_fed_52.txt 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064 0 2
## dispt_fed_53.txt 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040 0 5
## dispt_fed_54.txt 0.030 0.015 0.037 0.186 0.045 0.111 0.089 0.037 0 2

```

We will now add the results to the dataframe and visualize the clusters:

```

set.seed(22)
#first visualizing dataframe
head(fedpapers_km)

## a all also an and any are as at be
## dispt_fed_49.txt 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411
## dispt_fed_50.txt 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393
## dispt_fed_51.txt 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474
## dispt_fed_52.txt 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365
## dispt_fed_53.txt 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344
## dispt_fed_54.txt 0.245 0.059 0.007 0.067 0.282 0.052 0.111 0.252 0.015 0.297
## been but by can do down even every for. from
## dispt_fed_49.txt 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044
## dispt_fed_50.txt 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101
## dispt_fed_51.txt 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053
## dispt_fed_52.txt 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079

```



```

## dispt_fed_53.txt 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074
## dispt_fed_54.txt 0.030 0.037 0.186 0.000 0.000 0.007 0.007 0.007 0.067 0.096
##          had has have her his if. in. into is it
## dispt_fed_49.txt 0.035 0.017 0.044 0 0.017 0.000 0.262 0.009 0.157 0.175
## dispt_fed_50.txt 0.101 0.013 0.152 0 0.000 0.025 0.291 0.025 0.038 0.127
## dispt_fed_51.txt 0.008 0.015 0.023 0 0.000 0.023 0.308 0.038 0.150 0.173
## dispt_fed_52.txt 0.016 0.024 0.143 0 0.024 0.040 0.238 0.008 0.151 0.222
## dispt_fed_53.txt 0.000 0.054 0.047 0 0.020 0.034 0.263 0.013 0.189 0.108
## dispt_fed_54.txt 0.022 0.015 0.119 0 0.067 0.030 0.401 0.037 0.260 0.156
##          its may more must my no not now of on one
## dispt_fed_49.txt 0.070 0.035 0.026 0.026 0 0.035 0.114 0 0.900 0.140 0.026
## dispt_fed_50.txt 0.038 0.038 0.000 0.013 0 0.000 0.127 0 0.747 0.139 0.025
## dispt_fed_51.txt 0.030 0.120 0.038 0.083 0 0.030 0.068 0 0.858 0.150 0.030
## dispt_fed_52.txt 0.048 0.056 0.056 0.071 0 0.032 0.087 0 0.802 0.143 0.032
## dispt_fed_53.txt 0.013 0.047 0.067 0.013 0 0.047 0.128 0 0.869 0.054 0.047
## dispt_fed_54.txt 0.015 0.074 0.045 0.015 0 0.059 0.134 0 0.876 0.141 0.052
##          only or our shall should so some such than that
## dispt_fed_49.txt 0.035 0.096 0.017 0.017 0.017 0.035 0.009 0.026 0.009 0.184
## dispt_fed_50.txt 0.000 0.114 0.000 0.000 0.013 0.013 0.063 0.000 0.000 0.152
## dispt_fed_51.txt 0.023 0.060 0.000 0.008 0.068 0.038 0.030 0.045 0.023 0.188
## dispt_fed_52.txt 0.048 0.064 0.016 0.016 0.032 0.040 0.024 0.008 0.000 0.238
## dispt_fed_53.txt 0.027 0.081 0.027 0.000 0.000 0.027 0.067 0.027 0.047 0.162
## dispt_fed_54.txt 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
## dispt_fed_49.txt 1.425 0.114 0.000 0.009 0.009 0.044 0.507 0 0.000 0.009
## dispt_fed_50.txt 1.254 0.165 0.000 0.000 0.000 0.051 0.355 0 0.013 0.051
## dispt_fed_51.txt 1.490 0.053 0.015 0.015 0.000 0.075 0.361 0 0.000 0.008
## dispt_fed_52.txt 1.326 0.071 0.008 0.000 0.000 0.103 0.532 0 0.000 0.087
## dispt_fed_53.txt 1.193 0.027 0.007 0.007 0.000 0.094 0.485 0 0.000 0.027
## dispt_fed_54.txt 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 clusters
## dispt_fed_49.txt 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192 0 2
## dispt_fed_50.txt 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139 0 5

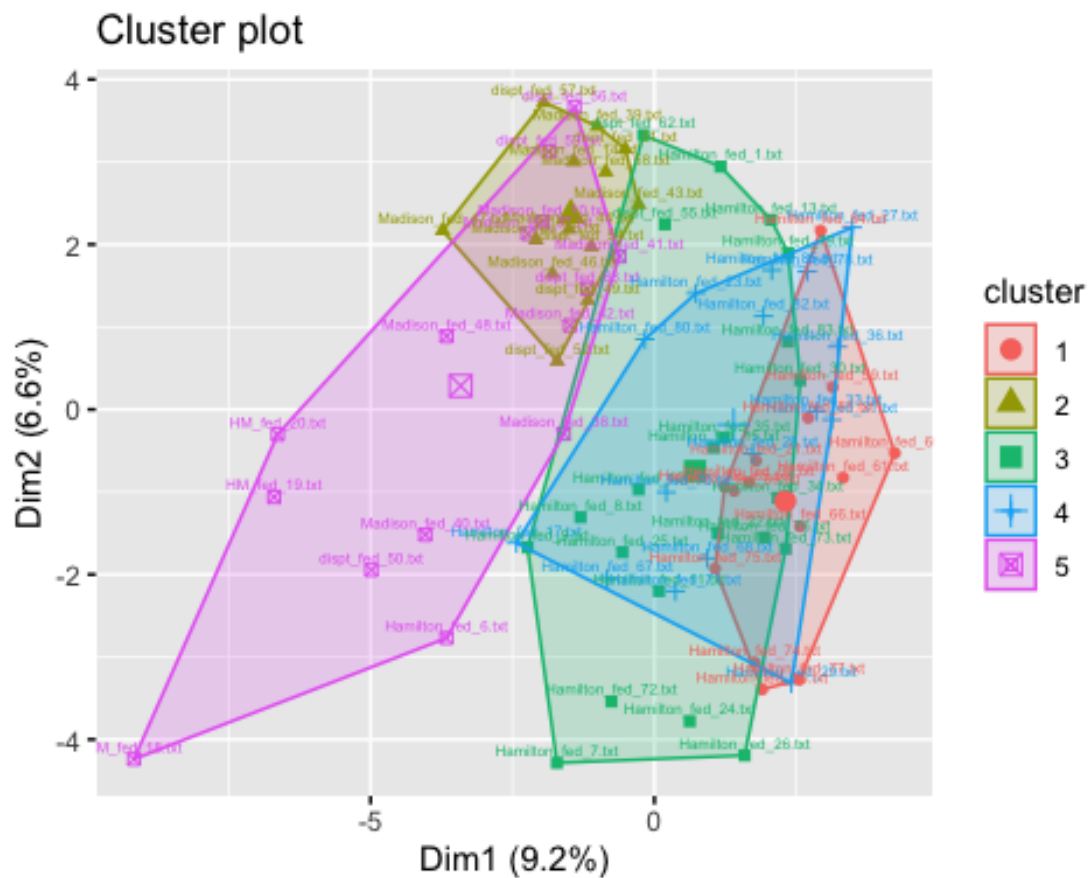
```

```
## dispt_fed_51.txt 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068 0 2
## dispt_fed_52.txt 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064 0 2
## dispt_fed_53.txt 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040 0 5
## dispt_fed_54.txt 0.030 0.015 0.037 0.186 0.045 0.111 0.089 0.037 0 2
```

#drop final column and visualize clusters

```
fedpapers_km_fviz<-fedpapers_km[, -71]
```

```
fviz_cluster(cluster, data=fedpapers_km_fviz, labels= 5)
```



Visualizing bar chart of various clusters

```
ggplot(data=fedpapers, aes(x=author, fill=clusters)) + geom_bar(stat="count") + labs(title = "Cluster Results from trial 1")
```



These cluster results tell us that the disputed papers are much more correlated with Madison's writing styles, which were shown in clusters 1 and 4. The disputed papers were predominantly shown in clusters 1 and 4 as well.

Hamilton's writing falls predominantly in clusters 3, 5, and slightly overlaps with cluster 1. Hamilton and Madison both overlap slightly in cluster 1. Based on the analysis of the clustering results, we chose a good initial number of clusters that showed a clear distinction between authors.

Based on these initial results, it seems much more likely that Madison is the author of these disputed papers.

Hierarchical Clustering Algorithms (HAC)

First, make a new df to pipe to the algorithm.

```
fedpapers_hac<-fedpapers[,2:72]
#making file name row names
rownames(fedpapers_hac)<-fedpapers_hac[,1]
fedpapers_hac[,1]<-NULL
```

#view df

head(fedpapers_hac)

```
##          a all also  an  and  any  are  as  at  be
## dispt_fed_49.txt 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411
## dispt_fed_50.txt 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393
## dispt_fed_51.txt 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474
## dispt_fed_52.txt 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365
## dispt_fed_53.txt 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344
## dispt_fed_54.txt 0.245 0.059 0.007 0.067 0.282 0.052 0.111 0.252 0.015 0.297
##          been but  by  can  do down even every for. from
## dispt_fed_49.txt 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044
## dispt_fed_50.txt 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101
## dispt_fed_51.txt 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053
## dispt_fed_52.txt 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079
## dispt_fed_53.txt 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074
## dispt_fed_54.txt 0.030 0.037 0.186 0.000 0.000 0.007 0.007 0.007 0.067 0.096
##          had  has have her  his  if.  in.  into  is  it
## dispt_fed_49.txt 0.035 0.017 0.044 0 0.017 0.000 0.262 0.009 0.157 0.175
## dispt_fed_50.txt 0.101 0.013 0.152 0 0.000 0.025 0.291 0.025 0.038 0.127
## dispt_fed_51.txt 0.008 0.015 0.023 0 0.000 0.023 0.308 0.038 0.150 0.173
## dispt_fed_52.txt 0.016 0.024 0.143 0 0.024 0.040 0.238 0.008 0.151 0.222
## dispt_fed_53.txt 0.000 0.054 0.047 0 0.020 0.034 0.263 0.013 0.189 0.108
## dispt_fed_54.txt 0.022 0.015 0.119 0 0.067 0.030 0.401 0.037 0.260 0.156
##          its  may more must my  no  not now  of  on  one
## dispt_fed_49.txt 0.070 0.035 0.026 0.026 0 0.035 0.114 0 0.900 0.140 0.026
## dispt_fed_50.txt 0.038 0.038 0.000 0.013 0 0.000 0.127 0 0.747 0.139 0.025
## dispt_fed_51.txt 0.030 0.120 0.038 0.083 0 0.030 0.068 0 0.858 0.150 0.030
## dispt_fed_52.txt 0.048 0.056 0.056 0.071 0 0.032 0.087 0 0.802 0.143 0.032
## dispt_fed_53.txt 0.013 0.047 0.067 0.013 0 0.047 0.128 0 0.869 0.054 0.047
## dispt_fed_54.txt 0.015 0.074 0.045 0.015 0 0.059 0.134 0 0.876 0.141 0.052
##          only  or  our shall should  so some such than that
## dispt_fed_49.txt 0.035 0.096 0.017 0.017 0.017 0.035 0.009 0.026 0.009 0.184
## dispt_fed_50.txt 0.000 0.114 0.000 0.000 0.013 0.013 0.063 0.000 0.000 0.152
```

```
## dispt_fed_51.txt 0.023 0.060 0.000 0.008 0.068 0.038 0.030 0.045 0.023 0.188
## dispt_fed_52.txt 0.048 0.064 0.016 0.016 0.032 0.040 0.024 0.008 0.000 0.238
## dispt_fed_53.txt 0.027 0.081 0.027 0.000 0.000 0.027 0.067 0.027 0.047 0.162
## dispt_fed_54.txt 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
## dispt_fed_49.txt 1.425 0.114 0.000 0.009 0.009 0.044 0.507 0 0.000 0.009
## dispt_fed_50.txt 1.254 0.165 0.000 0.000 0.000 0.051 0.355 0 0.013 0.051
## dispt_fed_51.txt 1.490 0.053 0.015 0.015 0.000 0.075 0.361 0 0.000 0.008
## dispt_fed_52.txt 1.326 0.071 0.008 0.000 0.000 0.103 0.532 0 0.000 0.087
## dispt_fed_53.txt 1.193 0.027 0.007 0.007 0.000 0.094 0.485 0 0.000 0.027
## dispt_fed_54.txt 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
## dispt_fed_49.txt 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192 0
## dispt_fed_50.txt 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139 0
## dispt_fed_51.txt 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068 0
## dispt_fed_52.txt 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064 0
## dispt_fed_53.txt 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040 0
## dispt_fed_54.txt 0.030 0.015 0.037 0.186 0.045 0.111 0.089 0.037 0
```

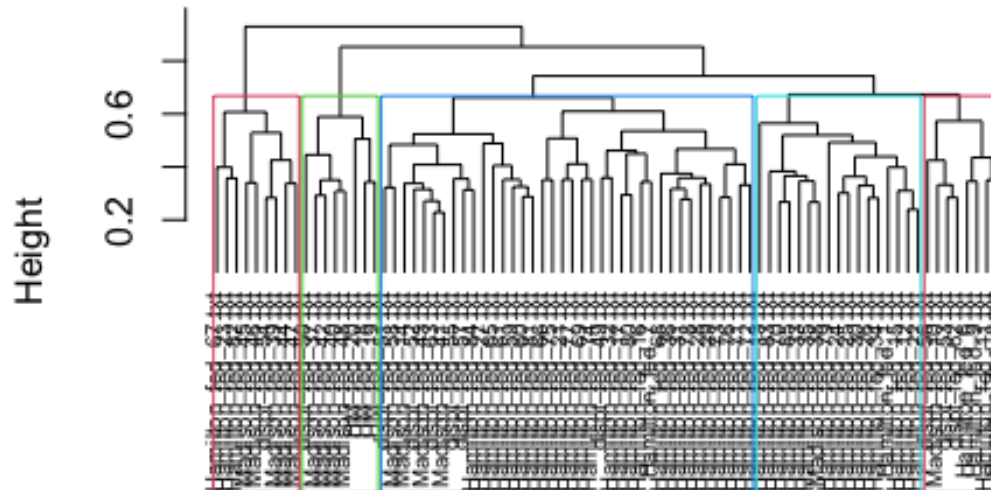
Now, we will calculate distance.

```
distance_euclidean<-dist(fedpapers_hac, method = "euclidean")
distance_maximum<-dist(fedpapers_hac, method = "maximum")
distance_manhattan<-dist(fedpapers_hac, method = "manhattan")
distance_canberra<-dist(fedpapers_hac, method = "canberra")
distance_binary<-dist(fedpapers_hac, method = "binary")
distance_minkowski<-dist(fedpapers_hac, method = "minkowski")
```

Cluster Dendrogram of Euclidean Distance

```
hac_euc<-hclust(distance_euclidean, method="complete")
plot(hac_euc, cex=0.6, hang=-1)
rect.hclust(hac_euc, k=5, border=2:5)
```

Cluster Dendrogram



distance_euclidean
hclust (*, "complete")

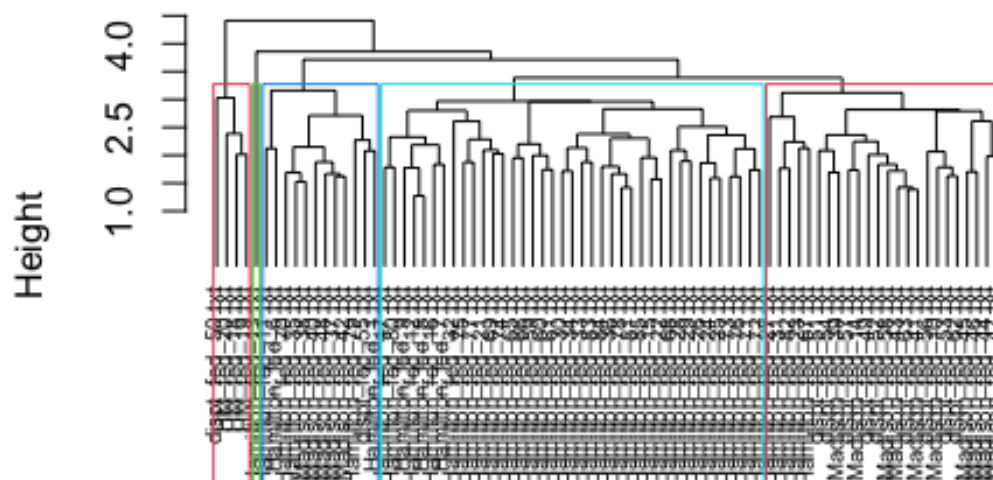
The

above diagram puts disputed papers in the leafs closest to Madison's papers as you read the diagram from the bottom up. This again suggests that the disputed papers are closest in similarity to Madison, and are most likely written by him.

Visualizing Dendrogram with Manhattan Distance

```
hac_man<-hclust(distance_manhattan, method="complete")
plot(hac_man, cex=0.6, hang=-1)
rect.hclust(hac_man, k=5, border=2:5)
```

Cluster Dendrogram



distance_manhattan
hclust (*, "complete")

The

Manhattan Distance Dendrogram is even more illustrative than the previous diagram, as it shows the large majority of disputed papers on the right handside. Here, we can clearly see that these disputed papers are closely connected to Madison.

Summary and Conclusions

To begin tackling the question of who wrote the disputed papers, we first began with some exploratory data analysis to see if there are clear differences in tone between the two authors. We found that Hamilton and Madison did in fact have a slight variance in tone (analyzing first/second person tone, as well as strong action words), with Hamilton being slightly more prone to use the first person and to use strong “call to action”. While this initial analysis did not help much with answering our final question, it did prove that the two writers had somewhat distinguishable writing styles before jumping into our algorithms.

First, we used the K-Means clustering algorithm to group the papers. We found that 5 clusters was a good K value, and provided clear categories that separated Madison and Hamilton- with Madison falling predominantly into clusters 1 and 4, while Hamilton fell into clusters 3, 5, and also in 1. Ultimately, Madison’s writing matched the disputed papers much more closely, supporting the conclusion that Madison wrote the papers.

When we conducted the HAC algorithm, we also found a strong correlation between Madison’s writing and the disputed papers. The Manhattan distance was more useful to

conclude this correlation, as it showed the disputed papers grouped together on the right handside which were closely related to Madison. Ultimately, this also supported the conclusion that Madison wrote the papers.

Finally, based on our K-Means algorithm, we were able to examine the papers written jointly by both Hamilton and Madison. These appeared to actually be in a cluster of their own (2), which only overlapped slightly with Hamilton's writing.