federalist-authorship\_unsupervised-learning

KA

2024-05-05

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

# Install required packages  
#install.packages(c("readr", "data.table", "cluster", "ggplot2", "factoextra"))  
  
# Load required libraries  
library(readr)  
library(data.table)  
library(cluster)  
library(ggplot2)  
library(factoextra)

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

# Load the dataset  
FederalistPapers <- read.csv("fedPapers85.csv")  
#Make a copy labelled original  
FederalistPapers\_original <- FederalistPapers  
# Check the structure of the dataset  
head(FederalistPapers)

## 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

str(FederalistPapers)

## 'data.frame': 85 obs. of 72 variables:  
## $ author : chr "dispt" "dispt" "dispt" "dispt" ...  
## $ filename: chr "dispt\_fed\_49.txt" "dispt\_fed\_50.txt" "dispt\_fed\_51.txt" "dispt\_fed\_52.txt" ...  
## $ a : num 0.28 0.177 0.339 0.27 0.303 0.245 0.349 0.414 0.248 0.442 ...  
## $ all : num 0.052 0.063 0.09 0.024 0.054 0.059 0.036 0.083 0.04 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 0.04 0.075 ...  
## $ and : num 0.358 0.393 0.301 0.262 0.404 0.282 0.335 0.478 0.356 0.423 ...  
## $ any : num 0.026 0.063 0.008 0.056 0.04 0.052 0.058 0.046 0.034 0.037 ...  
## $ are : num 0.131 0.051 0.068 0.064 0.128 0.111 0.087 0.11 0.154 0.093 ...  
## $ as : num 0.122 0.139 0.203 0.111 0.148 0.252 0.073 0.074 0.161 0.1 ...  
## $ at : num 0.017 0.114 0.023 0.056 0.013 0.015 0.116 0.037 0.047 0.031 ...  
## $ be : num 0.411 0.393 0.474 0.365 0.344 0.297 0.378 0.331 0.289 0.379 ...  
## $ been : num 0.026 0.165 0.015 0.127 0.047 0.03 0.044 0.046 0.027 0.025 ...  
## $ but : num 0.009 0 0.038 0.032 0.061 0.037 0.007 0.055 0.027 0.037 ...  
## $ by : num 0.14 0.139 0.173 0.167 0.209 0.186 0.102 0.092 0.168 0.174 ...  
## $ can : num 0.035 0 0.023 0.056 0.088 0 0.058 0.037 0.047 0.056 ...  
## $ do : num 0.026 0.013 0 0 0 0 0.015 0.028 0 0 ...  
## $ down : num 0 0 0.008 0 0 0.007 0 0 0 0 ...  
## $ even : num 0.009 0.025 0.015 0.024 0.02 0.007 0.007 0.018 0 0.006 ...  
## $ every : num 0.044 0 0.023 0.04 0.027 0.007 0.087 0.064 0.081 0.05 ...  
## $ for. : num 0.096 0.076 0.098 0.103 0.141 0.067 0.116 0.055 0.127 0.1 ...  
## $ from : num 0.044 0.101 0.053 0.079 0.074 0.096 0.08 0.083 0.074 0.124 ...  
## $ had : num 0.035 0.101 0.008 0.016 0 0.022 0.015 0.009 0.007 0 ...  
## $ has : num 0.017 0.013 0.015 0.024 0.054 0.015 0.036 0.037 0.02 0.019 ...  
## $ have : num 0.044 0.152 0.023 0.143 0.047 0.119 0.044 0.074 0.074 0.044 ...  
## $ her : num 0 0 0 0 0 0 0.007 0 0.034 0.025 ...  
## $ his : num 0.017 0 0 0.024 0.02 0.067 0 0.018 0.02 0.05 ...  
## $ if. : num 0 0.025 0.023 0.04 0.034 0.03 0.029 0 0 0.025 ...  
## $ in. : num 0.262 0.291 0.308 0.238 0.263 0.401 0.189 0.267 0.248 0.274 ...  
## $ into : num 0.009 0.025 0.038 0.008 0.013 0.037 0 0.037 0.013 0.037 ...  
## $ is : num 0.157 0.038 0.15 0.151 0.189 0.26 0.167 0.083 0.208 0.23 ...  
## $ it : num 0.175 0.127 0.173 0.222 0.108 0.156 0.102 0.165 0.134 0.131 ...  
## $ its : num 0.07 0.038 0.03 0.048 0.013 0.015 0 0.046 0.02 0.019 ...  
## $ may : num 0.035 0.038 0.12 0.056 0.047 0.074 0.08 0.092 0.027 0.106 ...  
## $ more : num 0.026 0 0.038 0.056 0.067 0.045 0.08 0.064 0.06 0.081 ...  
## $ must : num 0.026 0.013 0.083 0.071 0.013 0.015 0.044 0.018 0.027 0.068 ...  
## $ my : num 0 0 0 0 0 0 0.007 0 0 0 ...  
## $ no : num 0.035 0 0.03 0.032 0.047 0.059 0.022 0.018 0.02 0.044 ...  
## $ not : num 0.114 0.127 0.068 0.087 0.128 0.134 0.102 0.101 0.094 0.106 ...  
## $ now : num 0 0 0 0 0 0 0.007 0 0.007 0.012 ...  
## $ of : num 0.9 0.747 0.858 0.802 0.869 ...  
## $ on : num 0.14 0.139 0.15 0.143 0.054 0.141 0.051 0.083 0.127 0.118 ...  
## $ one : num 0.026 0.025 0.03 0.032 0.047 0.052 0.073 0.046 0.06 0.031 ...  
## $ only : num 0.035 0 0.023 0.048 0.027 0.022 0.007 0.046 0.02 0.012 ...  
## $ or : num 0.096 0.114 0.06 0.064 0.081 0.074 0.153 0.037 0.154 0.081 ...  
## $ our : num 0.017 0 0 0.016 0.027 0.03 0.051 0 0.007 0.025 ...  
## $ shall : num 0.017 0 0.008 0.016 0 0.015 0.007 0 0.02 0 ...  
## $ should : num 0.017 0.013 0.068 0.032 0 0.03 0.007 0 0 0.012 ...  
## $ so : num 0.035 0.013 0.038 0.04 0.027 0.007 0.051 0.018 0.04 0.05 ...  
## $ some : num 0.009 0.063 0.03 0.024 0.067 0.045 0.007 0.028 0.027 0.025 ...  
## $ such : num 0.026 0 0.045 0.008 0.027 0.015 0.015 0 0.013 0.031 ...  
## $ than : num 0.009 0 0.023 0 0.047 0.03 0.109 0.055 0.067 0.044 ...  
## $ that : num 0.184 0.152 0.188 0.238 0.162 0.208 0.233 0.165 0.208 0.218 ...  
## $ the : num 1.42 1.25 1.49 1.33 1.19 ...  
## $ their : num 0.114 0.165 0.053 0.071 0.027 0.089 0.109 0.083 0.154 0.081 ...  
## $ then : num 0 0 0.015 0.008 0.007 0.007 0.015 0.009 0.007 0.012 ...  
## $ there : num 0.009 0 0.015 0 0.007 0.007 0.036 0.028 0.02 0 ...  
## $ things : num 0.009 0 0 0 0 0 0 0 0 0.012 ...  
## $ this : num 0.044 0.051 0.075 0.103 0.094 0.126 0.08 0.11 0.067 0.093 ...  
## $ to : num 0.507 0.355 0.361 0.532 0.485 0.445 0.56 0.34 0.49 0.498 ...  
## $ up : num 0 0 0 0 0 0 0.007 0 0 0 ...  
## $ upon : num 0 0.013 0 0 0 0 0 0 0 0 ...  
## $ was : num 0.009 0.051 0.008 0.087 0.027 0.007 0.015 0.018 0.027 0 ...  
## $ were : num 0.017 0 0.015 0.079 0.02 0.03 0.029 0.009 0.007 0 ...  
## $ what : num 0 0 0.008 0.008 0.02 0.015 0.015 0.009 0.02 0.025 ...  
## $ when : num 0.009 0 0 0.024 0.007 0.037 0.007 0 0.02 0.012 ...  
## $ which : num 0.175 0.114 0.105 0.167 0.155 0.186 0.211 0.175 0.201 0.199 ...  
## $ who : num 0.044 0.038 0.008 0 0.027 0.045 0.022 0.018 0.04 0.031 ...  
## $ will : num 0.009 0.089 0.173 0.079 0.168 0.111 0.145 0.267 0.154 0.106 ...  
## $ with : num 0.087 0.063 0.045 0.079 0.074 0.089 0.073 0.129 0.027 0.081 ...  
## $ would : num 0.192 0.139 0.068 0.064 0.04 0.037 0.073 0.037 0.04 0.031 ...  
## $ your : num 0 0 0 0 0 0 0 0 0 0 ...

summary(FederalistPapers)

## author filename a all   
## Length:85 Length:85 Min. :0.0960 Min. :0.01500   
## Class :character Class :character 1st Qu.:0.2400 1st Qu.:0.03500   
## Mode :character Mode :character Median :0.2990 Median :0.05000   
## Mean :0.2932 Mean :0.05284   
## 3rd Qu.:0.3490 3rd Qu.:0.06600   
## Max. :0.4660 Max. :0.12700   
## also an and any   
## Min. :0.000000 Min. :0.00900 Min. :0.2170 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.04900 1st Qu.:0.3190 1st Qu.:0.02500   
## Median :0.007000 Median :0.07100 Median :0.3580 Median :0.04300   
## Mean :0.007659 Mean :0.06839 Mean :0.3846 Mean :0.04161   
## 3rd Qu.:0.013000 3rd Qu.:0.08500 3rd Qu.:0.4130 3rd Qu.:0.05600   
## Max. :0.047000 Max. :0.17900 Max. :0.8210 Max. :0.11400   
## are as at be   
## Min. :0.01300 Min. :0.0270 Min. :0.00000 Min. :0.0400   
## 1st Qu.:0.05100 1st Qu.:0.1000 1st Qu.:0.02600 1st Qu.:0.2580   
## Median :0.06800 Median :0.1240 Median :0.03800 Median :0.3070   
## Mean :0.07707 Mean :0.1242 Mean :0.04427 Mean :0.3012   
## 3rd Qu.:0.10200 3rd Qu.:0.1440 3rd Qu.:0.06300 3rd Qu.:0.3580   
## Max. :0.16300 Max. :0.2520 Max. :0.11800 Max. :0.4810   
## been but by can   
## Min. :0.00000 Min. :0.00000 Min. :0.0270 Min. :0.00000   
## 1st Qu.:0.03000 1st Qu.:0.02200 1st Qu.:0.0920 1st Qu.:0.01400   
## Median :0.05300 Median :0.03200 Median :0.1240 Median :0.02900   
## Mean :0.05967 Mean :0.03232 Mean :0.1272 Mean :0.03558   
## 3rd Qu.:0.08400 3rd Qu.:0.04200 3rd Qu.:0.1620 3rd Qu.:0.05200   
## Max. :0.16500 Max. :0.08900 Max. :0.2640 Max. :0.11000   
## do down even every   
## Min. :0.000000 Min. :0.000000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:0.00900   
## Median :0.006000 Median :0.000000 Median :0.0100 Median :0.02200   
## Mean :0.006259 Mean :0.001529 Mean :0.0114 Mean :0.02391   
## 3rd Qu.:0.010000 3rd Qu.:0.000000 3rd Qu.:0.0180 3rd Qu.:0.03400   
## Max. :0.028000 Max. :0.017000 Max. :0.0370 Max. :0.08700   
## for. from had has   
## Min. :0.03000 Min. :0.02600 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.07000 1st Qu.:0.05700 1st Qu.:0.00800 1st Qu.:0.02500   
## Median :0.08800 Median :0.07800 Median :0.01600 Median :0.04600   
## Mean :0.09376 Mean :0.07978 Mean :0.02116 Mean :0.04442   
## 3rd Qu.:0.11400 3rd Qu.:0.09800 3rd Qu.:0.02700 3rd Qu.:0.05700   
## Max. :0.21300 Max. :0.16200 Max. :0.14100 Max. :0.11400   
## have her his if.   
## Min. :0.01100 Min. :0.000000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.07300 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.01600   
## Median :0.09000 Median :0.000000 Median :0.01400 Median :0.02600   
## Mean :0.09474 Mean :0.008094 Mean :0.02862 Mean :0.02733   
## 3rd Qu.:0.12400 3rd Qu.:0.007000 3rd Qu.:0.03900 3rd Qu.:0.03400   
## Max. :0.18500 Max. :0.150000 Max. :0.24700 Max. :0.09900   
## in. into is it   
## Min. :0.1890 Min. :0.00000 Min. :0.0280 Min. :0.0750   
## 1st Qu.:0.2670 1st Qu.:0.01000 1st Qu.:0.1180 1st Qu.:0.1290   
## Median :0.3040 Median :0.02200 Median :0.1510 Median :0.1510   
## Mean :0.3174 Mean :0.02409 Mean :0.1563 Mean :0.1567   
## 3rd Qu.:0.3550 3rd Qu.:0.03400 3rd Qu.:0.1960 3rd Qu.:0.1900   
## Max. :0.4990 Max. :0.10500 Max. :0.3230 Max. :0.2840   
## its may more must   
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.03000 1st Qu.:0.03600 1st Qu.:0.02300 1st Qu.:0.01400   
## Median :0.04200 Median :0.05600 Median :0.04400 Median :0.02700   
## Mean :0.04836 Mean :0.06181 Mean :0.04561 Mean :0.03305   
## 3rd Qu.:0.06400 3rd Qu.:0.08500 3rd Qu.:0.06100 3rd Qu.:0.04400   
## Max. :0.15000 Max. :0.13400 Max. :0.13000 Max. :0.11100   
## my no not now   
## Min. :0.000000 Min. :0.00000 Min. :0.02000 Min. :0.000000   
## 1st Qu.:0.000000 1st Qu.:0.02000 1st Qu.:0.07500 1st Qu.:0.000000   
## Median :0.000000 Median :0.02900 Median :0.09500 Median :0.005000   
## Mean :0.003259 Mean :0.03236 Mean :0.09248 Mean :0.006035   
## 3rd Qu.:0.005000 3rd Qu.:0.04300 3rd Qu.:0.11200 3rd Qu.:0.010000   
## Max. :0.056000 Max. :0.08300 Max. :0.14800 Max. :0.026000   
## of on one only   
## Min. :0.5620 Min. :0.00000 Min. :0.00500 Min. :0.00000   
## 1st Qu.:0.8560 1st Qu.:0.04300 1st Qu.:0.02700 1st Qu.:0.01000   
## Median :0.9020 Median :0.06200 Median :0.03600 Median :0.02200   
## Mean :0.9094 Mean :0.06926 Mean :0.04079 Mean :0.02288   
## 3rd Qu.:0.9690 3rd Qu.:0.09700 3rd Qu.:0.05000 3rd Qu.:0.03400   
## Max. :1.2110 Max. :0.15600 Max. :0.13500 Max. :0.06500   
## or our shall should   
## Min. :0.02700 Min. :0.000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.07000 1st Qu.:0.000 1st Qu.:0.00600 1st Qu.:0.01000   
## Median :0.08100 Median :0.013 Median :0.01400 Median :0.02700   
## Mean :0.09674 Mean :0.023 Mean :0.01875 Mean :0.02656   
## 3rd Qu.:0.11600 3rd Qu.:0.028 3rd Qu.:0.02700 3rd Qu.:0.03800   
## Max. :0.32100 Max. :0.199 Max. :0.07900 Max. :0.09100   
## so some such than   
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.01800 1st Qu.:0.00900 1st Qu.:0.01800 1st Qu.:0.02700   
## Median :0.02900 Median :0.01700 Median :0.02900 Median :0.04300   
## Mean :0.02982 Mean :0.01989 Mean :0.02922 Mean :0.04396   
## 3rd Qu.:0.04000 3rd Qu.:0.02800 3rd Qu.:0.03800 3rd Qu.:0.05500   
## Max. :0.07200 Max. :0.06700 Max. :0.08500 Max. :0.15000   
## that the their then   
## Min. :0.081 Min. :0.669 Min. :0.00500 Min. :0.000000   
## 1st Qu.:0.171 1st Qu.:1.178 1st Qu.:0.05500 1st Qu.:0.000000   
## Median :0.208 Median :1.275 Median :0.08600 Median :0.006000   
## Mean :0.212 Mean :1.281 Mean :0.08553 Mean :0.006082   
## 3rd Qu.:0.244 3rd Qu.:1.423 3rd Qu.:0.10600 3rd Qu.:0.010000   
## Max. :0.380 Max. :1.803 Max. :0.18300 Max. :0.021000   
## there things this to   
## Min. :0.00000 Min. :0.000000 Min. :0.00900 Min. :0.3330   
## 1st Qu.:0.00900 1st Qu.:0.000000 1st Qu.:0.06900 1st Qu.:0.4690   
## Median :0.02200 Median :0.000000 Median :0.09000 Median :0.5400   
## Mean :0.02638 Mean :0.002659 Mean :0.08701 Mean :0.5358   
## 3rd Qu.:0.03900 3rd Qu.:0.006000 3rd Qu.:0.10500 3rd Qu.:0.6060   
## Max. :0.10500 Max. :0.015000 Max. :0.15300 Max. :0.7760   
## up upon was were   
## Min. :0.000000 Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.00900 1st Qu.:0.00700   
## Median :0.000000 Median :0.02800 Median :0.01500 Median :0.01500   
## Mean :0.003482 Mean :0.02922 Mean :0.02584 Mean :0.02022   
## 3rd Qu.:0.006000 3rd Qu.:0.05000 3rd Qu.:0.03200 3rd Qu.:0.02900   
## Max. :0.032000 Max. :0.10200 Max. :0.18900 Max. :0.10800   
## what when which who   
## Min. :0.00000 Min. :0.00000 Min. :0.0810 Min. :0.00000   
## 1st Qu.:0.00500 1st Qu.:0.00000 1st Qu.:0.1180 1st Qu.:0.01600   
## Median :0.01000 Median :0.00900 Median :0.1520 Median :0.02700   
## Mean :0.01286 Mean :0.01174 Mean :0.1578 Mean :0.03253   
## 3rd Qu.:0.02000 3rd Qu.:0.01500 3rd Qu.:0.1830 3rd Qu.:0.04400   
## Max. :0.06000 Max. :0.07300 Max. :0.2760 Max. :0.12900   
## will with would your   
## Min. :0.00600 Min. :0.02700 Min. :0.0090 Min. :0.000000   
## 1st Qu.:0.05200 1st Qu.:0.06100 1st Qu.:0.0420 1st Qu.:0.000000   
## Median :0.08100 Median :0.07900 Median :0.0780 Median :0.000000   
## Mean :0.09865 Mean :0.07968 Mean :0.1017 Mean :0.002024   
## 3rd Qu.:0.13500 3rd Qu.:0.09200 3rd Qu.:0.1470 3rd Qu.:0.000000   
## Max. :0.34000 Max. :0.15000 Max. :0.3820 Max. :0.074000

# Check for missing values  
sum(is.na(FederalistPapers))

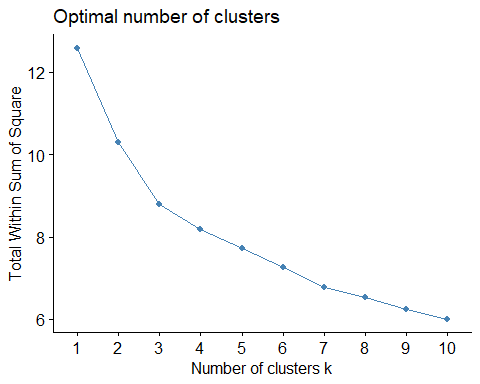
## [1] 0

# Remove columns that are unneccessary for clustering  
FedPapers\_km <-FederalistPapers[,2:72]  
 # Removed 'author'column  
head(FedPapers\_km)

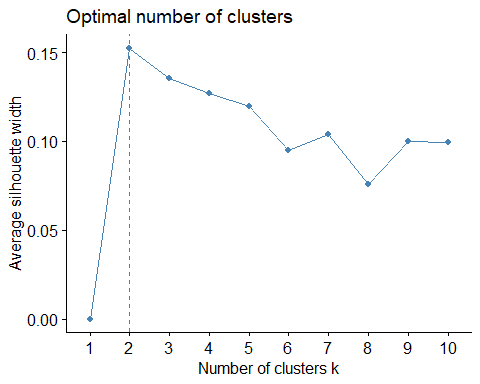
## filename a all also an and any are as at be  
## 1 dispt\_fed\_49.txt 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411  
## 2 dispt\_fed\_50.txt 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393  
## 3 dispt\_fed\_51.txt 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474  
## 4 dispt\_fed\_52.txt 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365  
## 5 dispt\_fed\_53.txt 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344  
## 6 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 had has have  
## 1 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044 0.035 0.017 0.044  
## 2 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101 0.101 0.013 0.152  
## 3 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053 0.008 0.015 0.023  
## 4 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079 0.016 0.024 0.143  
## 5 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074 0.000 0.054 0.047  
## 6 0.030 0.037 0.186 0.000 0.000 0.007 0.007 0.007 0.067 0.096 0.022 0.015 0.119  
## her his if. in. into is it its may more must my no  
## 1 0 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.026 0 0.035  
## 2 0 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.013 0 0.000  
## 3 0 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.083 0 0.030  
## 4 0 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.071 0 0.032  
## 5 0 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.013 0 0.047  
## 6 0 0.067 0.030 0.401 0.037 0.260 0.156 0.015 0.074 0.045 0.015 0 0.059  
## not now of on one only or our shall should so some such  
## 1 0.114 0 0.900 0.140 0.026 0.035 0.096 0.017 0.017 0.017 0.035 0.009 0.026  
## 2 0.127 0 0.747 0.139 0.025 0.000 0.114 0.000 0.000 0.013 0.013 0.063 0.000  
## 3 0.068 0 0.858 0.150 0.030 0.023 0.060 0.000 0.008 0.068 0.038 0.030 0.045  
## 4 0.087 0 0.802 0.143 0.032 0.048 0.064 0.016 0.016 0.032 0.040 0.024 0.008  
## 5 0.128 0 0.869 0.054 0.047 0.027 0.081 0.027 0.000 0.000 0.027 0.067 0.027  
## 6 0.134 0 0.876 0.141 0.052 0.022 0.074 0.030 0.015 0.030 0.007 0.045 0.015  
## than that the their then there things this to up upon was were  
## 1 0.009 0.184 1.425 0.114 0.000 0.009 0.009 0.044 0.507 0 0.000 0.009 0.017  
## 2 0.000 0.152 1.254 0.165 0.000 0.000 0.000 0.051 0.355 0 0.013 0.051 0.000  
## 3 0.023 0.188 1.490 0.053 0.015 0.015 0.000 0.075 0.361 0 0.000 0.008 0.015  
## 4 0.000 0.238 1.326 0.071 0.008 0.000 0.000 0.103 0.532 0 0.000 0.087 0.079  
## 5 0.047 0.162 1.193 0.027 0.007 0.007 0.000 0.094 0.485 0 0.000 0.027 0.020  
## 6 0.030 0.208 1.469 0.089 0.007 0.007 0.000 0.126 0.445 0 0.000 0.007 0.030  
## what when which who will with would your  
## 1 0.000 0.009 0.175 0.044 0.009 0.087 0.192 0  
## 2 0.000 0.000 0.114 0.038 0.089 0.063 0.139 0  
## 3 0.008 0.000 0.105 0.008 0.173 0.045 0.068 0  
## 4 0.008 0.024 0.167 0.000 0.079 0.079 0.064 0  
## 5 0.020 0.007 0.155 0.027 0.168 0.074 0.040 0  
## 6 0.015 0.037 0.186 0.045 0.111 0.089 0.037 0

# Make the file names the row names  
#rownames(FederalistPapers) <- FederalistPapers[, 1]  
  
# Remove the column containing the file names  
FederalistPapers <- FederalistPapers[, -1]  
  
View(FederalistPapers)  
# Now the dataframe is ready for k-means clustering

#library(factoextra)  
# Remove non-numeric columns (like filename or author)  
FedPapers\_num <- FederalistPapers[, sapply(FederalistPapers, is.numeric)]  
  
# Now use fviz\_nbclust on numeric-only data  
fviz\_nbclust(FedPapers\_num, kmeans, method = "wss")



fviz\_nbclust(FedPapers\_num, kmeans, method = "silhouette")



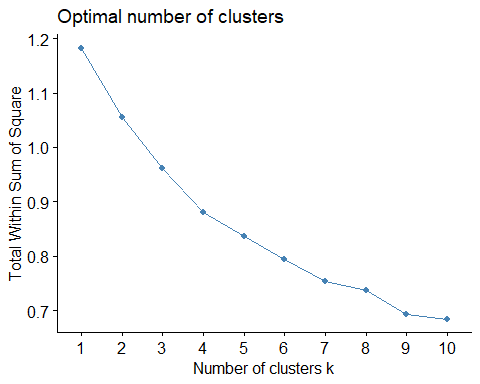
colnames(FederalistPapers)

## [1] "filename" "a" "all" "also" "an" "and"   
## [7] "any" "are" "as" "at" "be" "been"   
## [13] "but" "by" "can" "do" "down" "even"   
## [19] "every" "for." "from" "had" "has" "have"   
## [25] "her" "his" "if." "in." "into" "is"   
## [31] "it" "its" "may" "more" "must" "my"   
## [37] "no" "not" "now" "of" "on" "one"   
## [43] "only" "or" "our" "shall" "should" "so"   
## [49] "some" "such" "than" "that" "the" "their"   
## [55] "then" "there" "things" "this" "to" "up"   
## [61] "upon" "was" "were" "what" "when" "which"   
## [67] "who" "will" "with" "would" "your"

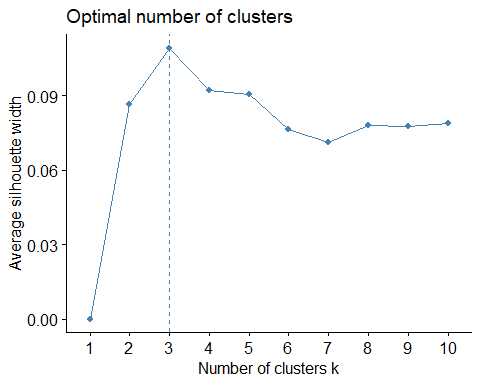
# Selecting only the desired columns  
FedPapers\_km <- FederalistPapers[, c("now", "up", "my", "things", "your", "down", "shall", "what", "when", "even", "her", "also", "do", "every", "our", "only", "had", "some", "but", "so", "such", "upon", "his", "if.", "should", "there", "than", "any", "one", "can", "must", "who", "no", "more", "all")]  
head(FedPapers\_km)

## now up my things your down shall what when even her also do every  
## 1 0 0 0 0.009 0 0.000 0.017 0.000 0.009 0.009 0 0.009 0.026 0.044  
## 2 0 0 0 0.000 0 0.000 0.000 0.000 0.000 0.025 0 0.013 0.013 0.000  
## 3 0 0 0 0.000 0 0.008 0.008 0.008 0.000 0.015 0 0.008 0.000 0.023  
## 4 0 0 0 0.000 0 0.000 0.016 0.008 0.024 0.024 0 0.016 0.000 0.040  
## 5 0 0 0 0.000 0 0.000 0.000 0.020 0.007 0.020 0 0.027 0.000 0.027  
## 6 0 0 0 0.000 0 0.007 0.015 0.015 0.037 0.007 0 0.007 0.000 0.007  
## our only had some but so such upon his if. should there  
## 1 0.017 0.035 0.035 0.009 0.009 0.035 0.026 0.000 0.017 0.000 0.017 0.009  
## 2 0.000 0.000 0.101 0.063 0.000 0.013 0.000 0.013 0.000 0.025 0.013 0.000  
## 3 0.000 0.023 0.008 0.030 0.038 0.038 0.045 0.000 0.000 0.023 0.068 0.015  
## 4 0.016 0.048 0.016 0.024 0.032 0.040 0.008 0.000 0.024 0.040 0.032 0.000  
## 5 0.027 0.027 0.000 0.067 0.061 0.027 0.027 0.000 0.020 0.034 0.000 0.007  
## 6 0.030 0.022 0.022 0.045 0.037 0.007 0.015 0.000 0.067 0.030 0.030 0.007  
## than any one can must who no more all  
## 1 0.009 0.026 0.026 0.035 0.026 0.044 0.035 0.026 0.052  
## 2 0.000 0.063 0.025 0.000 0.013 0.038 0.000 0.000 0.063  
## 3 0.023 0.008 0.030 0.023 0.083 0.008 0.030 0.038 0.090  
## 4 0.000 0.056 0.032 0.056 0.071 0.000 0.032 0.056 0.024  
## 5 0.047 0.040 0.047 0.088 0.013 0.027 0.047 0.067 0.054  
## 6 0.030 0.052 0.052 0.000 0.015 0.045 0.059 0.045 0.059

library(factoextra)  
  
# Using the elbow method  
fviz\_nbclust(FedPapers\_km, kmeans, method = "wss")



# Using the silhouette method  
fviz\_nbclust(FedPapers\_km, kmeans, method = "silhouette")

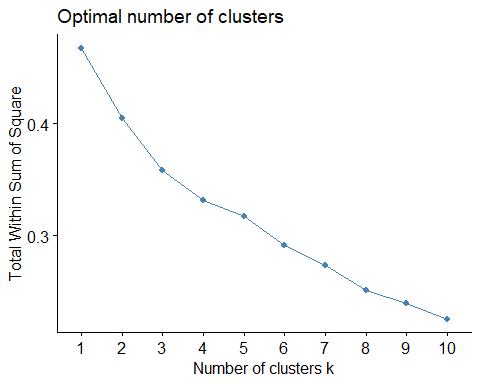


# Get the column sums  
word\_sums <- colSums(FedPapers\_km)  
  
# Order the column sums from highest to lowest  
word\_sums\_ordered <- sort(word\_sums, decreasing = TRUE)  
  
# Print the ordered column sums  
print(word\_sums\_ordered)

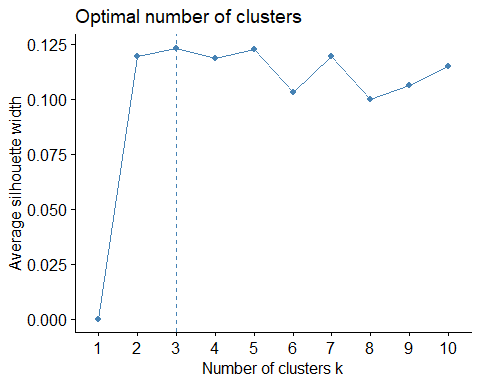
## all more than any one can must who no but so   
## 4.491 3.877 3.737 3.537 3.467 3.024 2.809 2.765 2.751 2.747 2.535   
## such upon his if. should there every our only had some   
## 2.484 2.484 2.433 2.323 2.258 2.242 2.032 1.955 1.945 1.799 1.691   
## shall what when even her also do now up my things   
## 1.594 1.093 0.998 0.969 0.688 0.651 0.532 0.513 0.296 0.277 0.226   
## your down   
## 0.172 0.130

#Revising selected columns  
FedPapers\_km <- FederalistPapers[, c("shall", "may", "even", "only", "upon","from", "also", "must", "all", "more", "should", "down", "up")]

library(factoextra)  
  
# Using the elbow method  
fviz\_nbclust(FedPapers\_km, kmeans, method = "wss")



# Using the silhouette method  
fviz\_nbclust(FedPapers\_km, kmeans, method = "silhouette")



set.seed(80) # Set seed for reproducibility  
k <- 10 # Number of clusters  
km\_model <- kmeans(FedPapers\_km, centers = k)

Clusters <- kmeans(FedPapers\_km, 10)  
FedPapers\_km$Clusters <- as.factor(Clusters$cluster)  
  
str(Clusters)

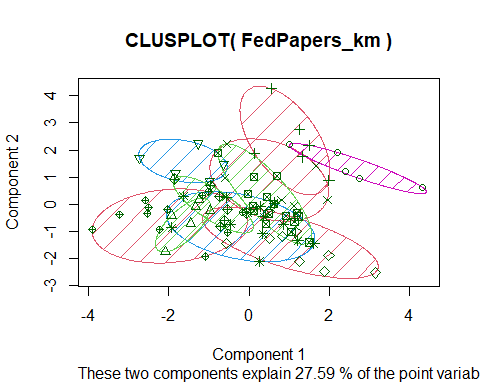
## List of 9  
## $ cluster : int [1:85] 7 9 6 6 4 6 6 6 7 6 ...  
## $ centers : num [1:10, 1:13] 0.0474 0.04417 0.02271 0.00633 0.0086 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:10] "1" "2" "3" "4" ...  
## .. ..$ : chr [1:13] "shall" "may" "even" "only" ...  
## $ totss : num 0.467  
## $ withinss : num [1:10] 0.01341 0.01863 0.02384 0.03066 0.00878 ...  
## $ tot.withinss: num 0.232  
## $ betweenss : num 0.235  
## $ size : int [1:10] 5 6 7 9 5 16 12 7 12 6  
## $ iter : int 4  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

Clusters$centers

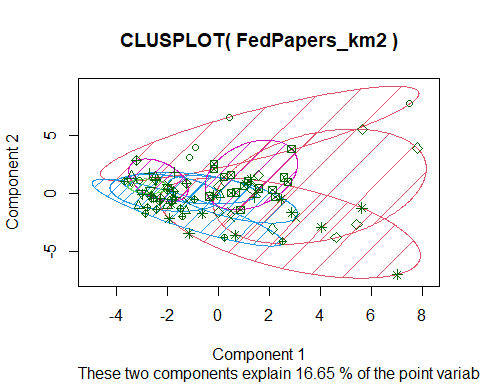
## shall may even only upon from  
## 1 0.047400000 0.10580000 0.003800000 0.01900000 0.02740000 0.08840000  
## 2 0.044166667 0.10383333 0.006333333 0.03233333 0.03650000 0.07300000  
## 3 0.022714286 0.05485714 0.008857143 0.01442857 0.01842857 0.05685714  
## 4 0.006333333 0.03688889 0.012444444 0.02233333 0.01200000 0.11577778  
## 5 0.008600000 0.07600000 0.005000000 0.01680000 0.08080000 0.06340000  
## 6 0.007812500 0.08612500 0.011562500 0.02981250 0.00181250 0.07425000  
## 7 0.026666667 0.02825000 0.011083333 0.02566667 0.01516667 0.06433333  
## 8 0.017142857 0.05300000 0.010000000 0.02071429 0.06314286 0.06671429  
## 9 0.014750000 0.04075000 0.019083333 0.02158333 0.05008333 0.08358333  
## 10 0.015166667 0.07150000 0.016000000 0.01350000 0.03883333 0.11900000  
## also must all more should down  
## 1 0.008200000 0.02340000 0.04300000 0.08260000 0.01080000 0.0000000000  
## 2 0.008833333 0.01916667 0.03766667 0.01700000 0.03016667 0.0016666667  
## 3 0.005428571 0.02257143 0.09942857 0.06171429 0.03085714 0.0017142857  
## 4 0.012000000 0.01966667 0.04688889 0.06544444 0.01222222 0.0000000000  
## 5 0.002200000 0.07640000 0.06780000 0.01700000 0.03080000 0.0000000000  
## 6 0.008812500 0.05006250 0.04837500 0.05506250 0.02793750 0.0011875000  
## 7 0.010833333 0.01900000 0.04275000 0.03533333 0.02208333 0.0008333333  
## 8 0.005571429 0.02271429 0.04342857 0.06185714 0.03085714 0.0054285714  
## 9 0.005416667 0.02900000 0.06475000 0.02591667 0.03458333 0.0015833333  
## 10 0.004166667 0.05400000 0.03750000 0.03450000 0.03333333 0.0036666667  
## up  
## 1 0.0030000000  
## 2 0.0013333333  
## 3 0.0050000000  
## 4 0.0023333333  
## 5 0.0048000000  
## 6 0.0017500000  
## 7 0.0043333333  
## 8 0.0021428571  
## 9 0.0078333333  
## 10 0.0006666667

FedPapers\_km2 <- FederalistPapers\_original  
FedPapers\_km2$Clusters <- as.factor(Clusters$cluster)

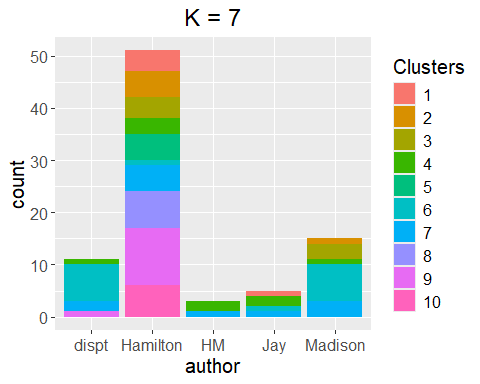
clusplot(FedPapers\_km, FedPapers\_km$Clusters, color = TRUE, shade = TRUE, labels = 0, lines = 0)



clusplot(FedPapers\_km2, FedPapers\_km2$Clusters, color = TRUE, shade = TRUE, labels = 0, lines = 0)



ggplot(data = FedPapers\_km2, aes(x= author, fill=Clusters))+  
 geom\_bar(stat = "count")+  
 labs(title = "K = 7")+  
 theme(plot.title = element\_text(hjust = 0.5), text=element\_text(size = 15))



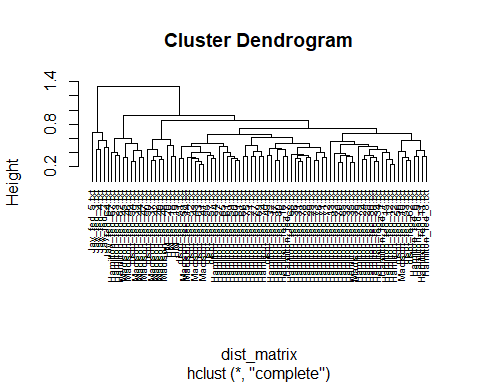
#It Appears that Madison and the disputed papers have the most highest and matching count in cluster 6.

FedPapers\_HAC <- FederalistPapers\_original[,c(2:72)]  
##Make the file names the row names. Need a data frame of numerical values for HAC  
rownames(FedPapers\_HAC) <- FedPapers\_HAC[,1]  
  
FedPapers\_HAC[,1] <- NULL

Interpretation: The K-means clustering (k = 10) grouped essays based on function word usage. When visualizing clusters, we observe that Madison’s essays and the disputed papers are often grouped in the same cluster (notably cluster 6), suggesting stylistic similarity. Hamilton’s essays appear more broadly distributed across clusters, while Jay’s and co-authored essays are more distinct.

Insight: Cluster assignment patterns align with historical evidence that James Madison is the more likely author of the disputed essays. This aligns with Mosteller and Wallace’s original findings.

# Compute the distance matrix  
dist\_matrix <- dist(FedPapers\_HAC)  
#dist\_matrix (commented out to prevent overwhelming output)  
  
# Perform hierarchical clustering  
HAC <- hclust(dist\_matrix, method = "complete")  
  
# Plot the dendrogram with smaller font size for labels  
plot(HAC, hang = -1, cex = 0.6)

 Interpretation: The hierarchical clustering dendrograms (complete and single linkage) show that disputed essays are often closer to Madison’s essays in the tree. This further supports the finding from K-means.

##Calculate distance using different methods  
distance <- dist(FedPapers\_HAC, method = "euclidean")  
distance2 <- dist(FedPapers\_HAC, method = "maximum")  
distance3 <- dist(FedPapers\_HAC, method = "manhattan")  
distance4 <- dist(FedPapers\_HAC, method = "canberra")  
distance5 <- dist(FedPapers\_HAC, method = "binary")  
distance6 <- dist(FedPapers\_HAC, method = "minkowski")

summary(distance)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2284 0.4187 0.4896 0.5241 0.5871 1.3339

summary(distance2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0800 0.1960 0.2530 0.2835 0.3350 1.1340

summary(distance3)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.273 2.234 2.523 2.624 2.887 4.871

summary(distance4)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.51 23.77 26.21 26.53 29.08 38.79

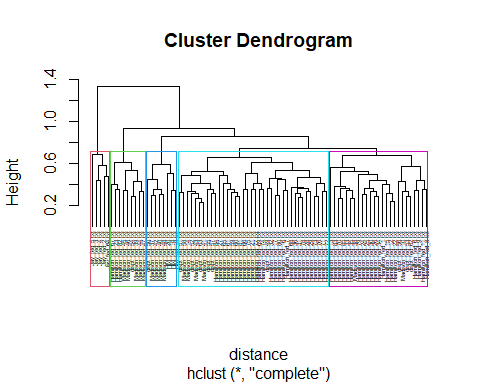
summary(distance5)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1194 0.1515 0.1540 0.1846 0.3538

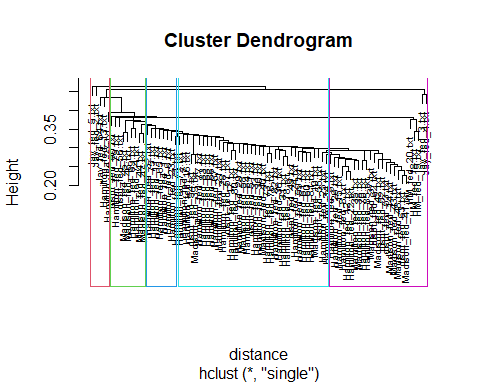
summary(distance6)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2284 0.4187 0.4896 0.5241 0.5871 1.3339

HAC <- hclust(distance, method = "complete")  
plot(HAC, cex=0.4, hang=-1)  
rect.hclust(HAC, k=5, border=2:9)



HAC2 <- hclust(distance, method = "single")  
plot(HAC2, cex=0.6)  
rect.hclust(HAC, k=5, border=2:8)

 Conclusion: Unsupervised learning methods like K-means and hierarchical clustering proved valuable in exploring historical text authorship. While not definitive, the patterns observed — including cluster groupings and proximity in dendrograms — support the hypothesis that Madison is the more likely author of the disputed Federalist Papers.