HW4 - Clustering

Using K-Means and HAC, you are going to try solving this mystery using clustering algorithms. Document your analysis process and draw your conclusion on who wrote the disputed essays. Provide evidence for each method to demonstrate what patterns had been learned to predict the disputed papers, for example, visualize the clustering results and show where the disputed papers are located in relation to Hamilton and Madison's papers. By the way, where are the papers with joint authorship located? For k-Means, analyze the centroids to explain which attributes are most useful for clustering. Hint: the centroid values on these dimensions should be far apart from each other to be able to distinguish the clusters.

```
#Name: Hrishikesh Telang
```

I am loading the required packages

```
library(factoextra)
library(stringr)
library(rpart)
library(caret)
library(gridExtra)
library(tidyr)
```

Now I am loading the dataset

```
df <- read.csv("HW4-data-fedPapers85.csv")
```

I am trying to check the structure of the dataframe

```
str(df)
```

```
## '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  1.43 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 ...
```

I am trying to check the structure of the dataframe

```
View(df)
```

Data Manipulation to label the data points:

I am creating a new column with a short form of the author name and storing it as modified_author:

```
df$modified_author <- ifelse(df$author == 'dispt', 'D', ifelse(df$author == 'Hamilton', 'H', ifelse(df$a
df$modified_author
```

```
##  [1] "D"  "D"  "D"  "D"  "D"  "D"  "D"  "D"  "D"  "D"  "D"  "H"  "H"  "H"  "H"
## [16] "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"
## [31] "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"
## [46] "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"  "H"
## [61] "H"  "H"  "HM" "HM" "HM" "J"  "J"  "J"  "J"  "J"  "M"  "M"  "M"  "M"  "M"
## [76] "M"  "M"  "M"  "M"  "M"  "M"  "M"  "M"  "M"  "M"
```

I am splitting the file name and number into Name and Num:

```
df <- extract(df, filename, into = c("Name", "Num"), "([^(]+)\\s*[^0-9]+([0-9].).")
```

I am creating a new column combining the author name along with the file number:

```
df$file <- paste(df$modified_author, "-", df$Num)
```

I convert the column to index:

```
rownames(df)<-df$file
head(df, 5)
```

```
##        author      Name Num     a   all  also    an   and   any   are    as
## D - 49  dispt dispt_fed  49 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122
## D - 50  dispt dispt_fed  50 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139
## D - 51  dispt dispt_fed  51 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203
## D - 52  dispt dispt_fed  52 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111
## D - 53  dispt dispt_fed  53 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148
```

```
##           at    be  been   but    by   can    do  down  even every  for.  from
## D - 49 0.017 0.411 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044
## D - 50 0.114 0.393 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101
## D - 51 0.023 0.474 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053
## D - 52 0.056 0.365 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079
## D - 53 0.013 0.344 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074
##          had   has  have her   his   if.   in.  into    is    it   its   may
## D - 49 0.035 0.017 0.044   0 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035
## D - 50 0.101 0.013 0.152   0 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038
## D - 51 0.008 0.015 0.023   0 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120
## D - 52 0.016 0.024 0.143   0 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056
## D - 53 0.000 0.054 0.047   0 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047
##         more  must my    no   not now    of    on   one  only    or   our shall
## D - 49 0.026 0.026  0 0.035 0.114   0 0.900 0.140 0.026 0.035 0.096 0.017 0.017
## D - 50 0.000 0.013  0 0.000 0.127   0 0.747 0.139 0.025 0.000 0.114 0.000 0.000
## D - 51 0.038 0.083  0 0.030 0.068   0 0.858 0.150 0.030 0.023 0.060 0.000 0.008
## D - 52 0.056 0.071  0 0.032 0.087   0 0.802 0.143 0.032 0.048 0.064 0.016 0.016
## D - 53 0.067 0.013  0 0.047 0.128   0 0.869 0.054 0.047 0.027 0.081 0.027 0.000
##        should    so  some  such  than  that   the their  then there things
## D - 49  0.017 0.035 0.009 0.026 0.009 0.184 1.425 0.114 0.000 0.009  0.009
## D - 50  0.013 0.013 0.063 0.000 0.000 0.152 1.254 0.165 0.000 0.000  0.000
## D - 51  0.068 0.038 0.030 0.045 0.023 0.188 1.490 0.053 0.015 0.015  0.000
## D - 52  0.032 0.040 0.024 0.008 0.000 0.238 1.326 0.071 0.008 0.000  0.000
## D - 53  0.000 0.027 0.067 0.027 0.047 0.162 1.193 0.027 0.007 0.007  0.000
##         this    to up  upon   was  were  what  when which   who  will  with
## D - 49 0.044 0.507  0 0.000 0.009 0.017 0.000 0.009 0.175 0.044 0.009 0.087
## D - 50 0.051 0.355  0 0.013 0.051 0.000 0.000 0.000 0.114 0.038 0.089 0.063
## D - 51 0.075 0.361  0 0.000 0.008 0.015 0.008 0.000 0.105 0.008 0.173 0.045
## D - 52 0.103 0.532  0 0.000 0.087 0.079 0.008 0.024 0.167 0.000 0.079 0.079
## D - 53 0.094 0.485  0 0.000 0.027 0.020 0.020 0.007 0.155 0.027 0.168 0.074
##        would your modified_author   file
## D - 49 0.192    0               D D - 49
## D - 50 0.139    0               D D - 50
## D - 51 0.068    0               D D - 51
## D - 52 0.064    0               D D - 52
## D - 53 0.040    0               D D - 53
```

I remove all the useless columns

```
df <- df[-c(1,2,3)]
df <- df[c(-(ncol(df)))]
head(df, 5)
```

```
##           a   all  also    an   and   any   are    as    at    be  been   but
## D - 49 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411 0.026 0.009
## D - 50 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393 0.165 0.000
## D - 51 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474 0.015 0.038
## D - 52 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365 0.127 0.032
## D - 53 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344 0.047 0.061
##           by   can    do  down  even every  for.  from   had   has  have her
## D - 49 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044 0.035 0.017 0.044   0
## D - 50 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101 0.101 0.013 0.152   0
```

```
## D - 51 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053 0.008 0.015 0.023   0
## D - 52 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079 0.016 0.024 0.143   0
## D - 53 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074 0.000 0.054 0.047   0
##          his   if.   in.  into    is    it   its   may  more  must my    no
## D - 49 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.026  0 0.035
## D - 50 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.013  0 0.000
## D - 51 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.083  0 0.030
## D - 52 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.071  0 0.032
## D - 53 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.013  0 0.047
##          not now    of    on   one  only    or   our shall should    so  some
## D - 49 0.114   0 0.900 0.140 0.026 0.035 0.096 0.017 0.017  0.017 0.035 0.009
## D - 50 0.127   0 0.747 0.139 0.025 0.000 0.114 0.000 0.000  0.013 0.013 0.063
## D - 51 0.068   0 0.858 0.150 0.030 0.023 0.060 0.000 0.008  0.068 0.038 0.030
## D - 52 0.087   0 0.802 0.143 0.032 0.048 0.064 0.016 0.016  0.032 0.040 0.024
## D - 53 0.128   0 0.869 0.054 0.047 0.027 0.081 0.027 0.000  0.000 0.027 0.067
##         such  than  that   the their  then there things  this    to up  upon
## D - 49 0.026 0.009 0.184 1.425 0.114 0.000 0.009  0.009 0.044 0.507  0 0.000
## D - 50 0.000 0.000 0.152 1.254 0.165 0.000 0.000  0.000 0.051 0.355  0 0.013
## D - 51 0.045 0.023 0.188 1.490 0.053 0.015 0.015  0.000 0.075 0.361  0 0.000
## D - 52 0.008 0.000 0.238 1.326 0.071 0.008 0.000  0.000 0.103 0.532  0 0.000
## D - 53 0.027 0.047 0.162 1.193 0.027 0.007 0.007  0.000 0.094 0.485  0 0.000
##          was  were  what  when which   who  will  with would your
## D - 49 0.009 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192    0
## D - 50 0.051 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139    0
## D - 51 0.008 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068    0
## D - 52 0.087 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064    0
## D - 53 0.027 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040    0
##        modified_author
## D - 49               D
## D - 50               D
## D - 51               D
## D - 52               D
## D - 53               D
```

Dropping the rows of files authored by Jay and Hamilton+Madison:

As we are only concerned about the authorship of the disputed articles, of Hamilton and of Madison, we are not concerned about those 3 articles written by Hamilton and Madison and 5 written by Jay. Thus, we can go ahead and remove 'Jay' and 'HM' from the dataframe and store it in the dataframe 'subset'.

```
subset <- subset(df, df$modified_author!="J" & df$modified_author!="HM")
head(subset, 5)
```

```
##            a   all  also    an   and   any   are    as    at    be  been   but
## D - 49 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411 0.026 0.009
## D - 50 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393 0.165 0.000
## D - 51 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474 0.015 0.038
## D - 52 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365 0.127 0.032
## D - 53 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344 0.047 0.061
```

```
##           by   can    do  down  even every  for.  from   had   has  have her
## D - 49 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044 0.035 0.017 0.044   0
## D - 50 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101 0.101 0.013 0.152   0
## D - 51 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053 0.008 0.015 0.023   0
## D - 52 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079 0.016 0.024 0.143   0
## D - 53 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074 0.000 0.054 0.047   0
##          his   if.   in.  into    is    it   its   may  more  must my    no
## D - 49 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.026  0 0.035
## D - 50 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.013  0 0.000
## D - 51 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.083  0 0.030
## D - 52 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.071  0 0.032
## D - 53 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.013  0 0.047
##          not now    of    on   one  only    or   our shall should    so  some
## D - 49 0.114   0 0.900 0.140 0.026 0.035 0.096 0.017 0.017  0.017 0.035 0.009
## D - 50 0.127   0 0.747 0.139 0.025 0.000 0.114 0.000 0.000  0.013 0.013 0.063
## D - 51 0.068   0 0.858 0.150 0.030 0.023 0.060 0.000 0.008  0.068 0.038 0.030
## D - 52 0.087   0 0.802 0.143 0.032 0.048 0.064 0.016 0.016  0.032 0.040 0.024
## D - 53 0.128   0 0.869 0.054 0.047 0.027 0.081 0.027 0.000  0.000 0.027 0.067
##         such  than  that   the their  then there things  this    to up  upon
## D - 49 0.026 0.009 0.184 1.425 0.114 0.000 0.009  0.009 0.044 0.507  0 0.000
## D - 50 0.000 0.000 0.152 1.254 0.165 0.000 0.000  0.000 0.051 0.355  0 0.013
## D - 51 0.045 0.023 0.188 1.490 0.053 0.015 0.015  0.000 0.075 0.361  0 0.000
## D - 52 0.008 0.000 0.238 1.326 0.071 0.008 0.000  0.000 0.103 0.532  0 0.000
## D - 53 0.027 0.047 0.162 1.193 0.027 0.007 0.007  0.000 0.094 0.485  0 0.000
##          was  were  what  when which   who  will  with would your
## D - 49 0.009 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192    0
## D - 50 0.051 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139    0
## D - 51 0.008 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068    0
## D - 52 0.087 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064    0
## D - 53 0.027 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040    0
##        modified_author
## D - 49               D
## D - 50               D
## D - 51               D
## D - 52               D
## D - 53               D
```

Dropping unused levels:

```
subset <- droplevels(subset)
```

I am just checking the first five rows of subset:

```
head(subset, 5)
```

```
##           a   all  also    an   and   any   are    as    at    be  been   but
## D - 49 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411 0.026 0.009
## D - 50 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393 0.165 0.000
## D - 51 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474 0.015 0.038
```

```
## D - 52 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365 0.127 0.032
## D - 53 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344 0.047 0.061
##           by   can    do  down  even every  for.  from   had   has  have her
## D - 49 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044 0.035 0.017 0.044   0
## D - 50 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101 0.101 0.013 0.152   0
## D - 51 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053 0.008 0.015 0.023   0
## D - 52 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079 0.016 0.024 0.143   0
## D - 53 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074 0.000 0.054 0.047   0
##          his   if.   in.  into    is    it   its   may  more  must my    no
## D - 49 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.026  0 0.035
## D - 50 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.013  0 0.000
## D - 51 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.083  0 0.030
## D - 52 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.071  0 0.032
## D - 53 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.013  0 0.047
##          not now    of    on   one  only    or   our shall should    so  some
## D - 49 0.114   0 0.900 0.140 0.026 0.035 0.096 0.017 0.017  0.017 0.035 0.009
## D - 50 0.127   0 0.747 0.139 0.025 0.000 0.114 0.000 0.000  0.013 0.013 0.063
## D - 51 0.068   0 0.858 0.150 0.030 0.023 0.060 0.000 0.008  0.068 0.038 0.030
## D - 52 0.087   0 0.802 0.143 0.032 0.048 0.064 0.016 0.016  0.032 0.040 0.024
## D - 53 0.128   0 0.869 0.054 0.047 0.027 0.081 0.027 0.000  0.000 0.027 0.067
##         such  than  that   the their  then there things  this    to up  upon
## D - 49 0.026 0.009 0.184 1.425 0.114 0.000 0.009  0.009 0.044 0.507  0 0.000
## D - 50 0.000 0.000 0.152 1.254 0.165 0.000 0.000  0.000 0.051 0.355  0 0.013
## D - 51 0.045 0.023 0.188 1.490 0.053 0.015 0.015  0.000 0.075 0.361  0 0.000
## D - 52 0.008 0.000 0.238 1.326 0.071 0.008 0.000  0.000 0.103 0.532  0 0.000
## D - 53 0.027 0.047 0.162 1.193 0.027 0.007 0.007  0.000 0.094 0.485  0 0.000
##          was  were  what  when which   who  will  with would your
## D - 49 0.009 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192    0
## D - 50 0.051 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139    0
## D - 51 0.008 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068    0
## D - 52 0.087 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064    0
## D - 53 0.027 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040    0
##        modified_author
## D - 49               D
## D - 50               D
## D - 51               D
## D - 52               D
## D - 53               D
```

I create a copy of subset2:

```
subset2 <- data.frame(subset)
head(subset2, 5)
```

```
##             a   all  also    an   and   any   are    as    at    be  been   but
## D - 49 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122 0.017 0.411 0.026 0.009
## D - 50 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139 0.114 0.393 0.165 0.000
## D - 51 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203 0.023 0.474 0.015 0.038
## D - 52 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111 0.056 0.365 0.127 0.032
## D - 53 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148 0.013 0.344 0.047 0.061
##           by   can    do  down  even every  for.  from   had   has  have her
## D - 49 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044 0.035 0.017 0.044   0
```

```
## D - 50 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101 0.101 0.013 0.152   0
## D - 51 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053 0.008 0.015 0.023   0
## D - 52 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079 0.016 0.024 0.143   0
## D - 53 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074 0.000 0.054 0.047   0
##          his   if.   in.  into    is    it   its   may  more  must my    no
## D - 49 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035 0.026 0.026  0 0.035
## D - 50 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038 0.000 0.013  0 0.000
## D - 51 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120 0.038 0.083  0 0.030
## D - 52 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056 0.056 0.071  0 0.032
## D - 53 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047 0.067 0.013  0 0.047
##          not now    of    on   one  only    or   our shall should    so  some
## D - 49 0.114   0 0.900 0.140 0.026 0.035 0.096 0.017 0.017  0.017 0.035 0.009
## D - 50 0.127   0 0.747 0.139 0.025 0.000 0.114 0.000 0.000  0.013 0.013 0.063
## D - 51 0.068   0 0.858 0.150 0.030 0.023 0.060 0.000 0.008  0.068 0.038 0.030
## D - 52 0.087   0 0.802 0.143 0.032 0.048 0.064 0.016 0.016  0.032 0.040 0.024
## D - 53 0.128   0 0.869 0.054 0.047 0.027 0.081 0.027 0.000  0.000 0.027 0.067
##         such  than  that   the their  then there things  this    to up  upon
## D - 49 0.026 0.009 0.184 1.425 0.114 0.000 0.009  0.009 0.044 0.507  0 0.000
## D - 50 0.000 0.000 0.152 1.254 0.165 0.000 0.000  0.000 0.051 0.355  0 0.013
## D - 51 0.045 0.023 0.188 1.490 0.053 0.015 0.015  0.000 0.075 0.361  0 0.000
## D - 52 0.008 0.000 0.238 1.326 0.071 0.008 0.000  0.000 0.103 0.532  0 0.000
## D - 53 0.027 0.047 0.162 1.193 0.027 0.007 0.007  0.000 0.094 0.485  0 0.000
##          was  were  what  when which   who  will  with would your
## D - 49 0.009 0.017 0.000 0.009 0.175 0.044 0.009 0.087 0.192    0
## D - 50 0.051 0.000 0.000 0.000 0.114 0.038 0.089 0.063 0.139    0
## D - 51 0.008 0.015 0.008 0.000 0.105 0.008 0.173 0.045 0.068    0
## D - 52 0.087 0.079 0.008 0.024 0.167 0.000 0.079 0.079 0.064    0
## D - 53 0.027 0.020 0.020 0.007 0.155 0.027 0.168 0.074 0.040    0
##        modified_author
## D - 49               D
## D - 50               D
## D - 51               D
## D - 52               D
## D - 53               D
```

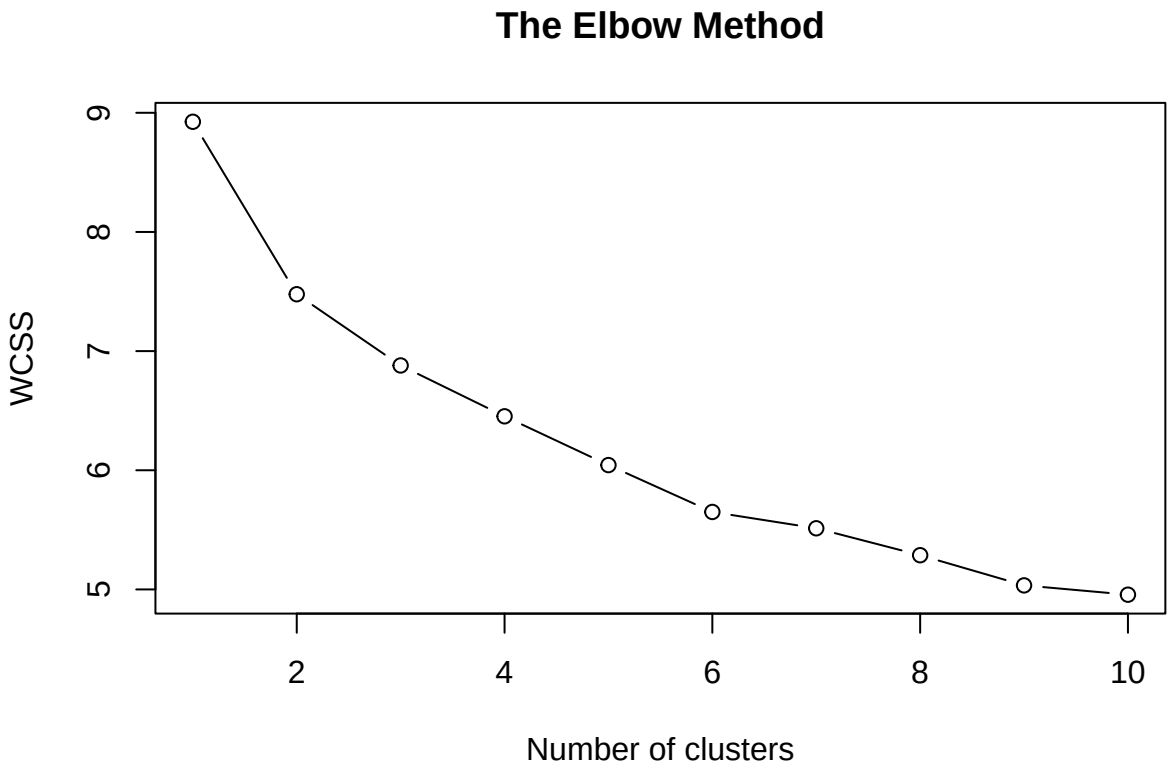
K-means - Default

Clustering is an unsupervised learning technique. It is the task of grouping together a set of objects in a way that objects in the same cluster are more similar to each other than to objects in other clusters. Similarity is an amount that reflects the strength of relationship between two data objects. Clustering is mainly used for exploratory data mining. It is used in many fields such as machine learning, pattern recognition, image analysis, information retrieval, bio-informatics, data compression, and computer graphics.

I use the elbow method directly here. The elbow method helps us check the optimality of the clusters required for analysis:

```
set.seed(6)
wcss = vector()
for (i in 1:10) wcss[i] = sum(kmeans(subset2[1:(length(subset2)-1)], i)$withinss)
plot(1:10,
     wcss,
```

```
     type = 'b',
     main = paste('The Elbow Method'),
     xlab = 'Number of clusters',
     ylab = 'WCSS')
```



From the above graph, it is safe to say that 5 or 6 are the optimal number of clusters for this dataset. For this example, let us consider 5 clusters

We train the K Means algorithm by removing the last column and taking 5 clusters:

```
set.seed(29)
kmeans = kmeans(x = subset2[1:(length(subset2)-1)], centers = 5)
y_kmeans = kmeans$cluster
```

We check how many disputed articles were linked with which author:

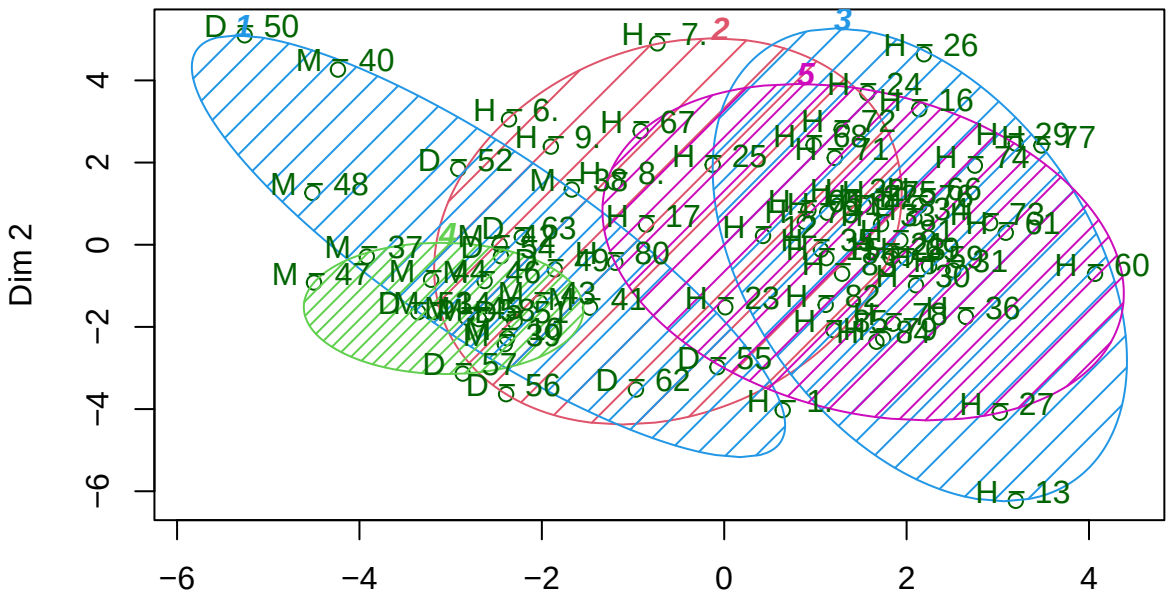
```
t <- t(table(subset[,length(subset)], y_kmeans))
t
```

```
##
## y_kmeans  D  H  M
##        1  6  1  8
##        2  1  9  0
##        3  0 18  0
##        4  4  0  7
##        5  0 23  0
```

We can clearly see that the disputed articles are authored by Madison as there is a strong link of similarity there.

Visualising the clusters

```
library(cluster)
clusplot(subset,
         y_kmeans,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels = 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste('Clusters of customers'),
         xlab = 'Dim 1',
         ylab = 'Dim 2')
```

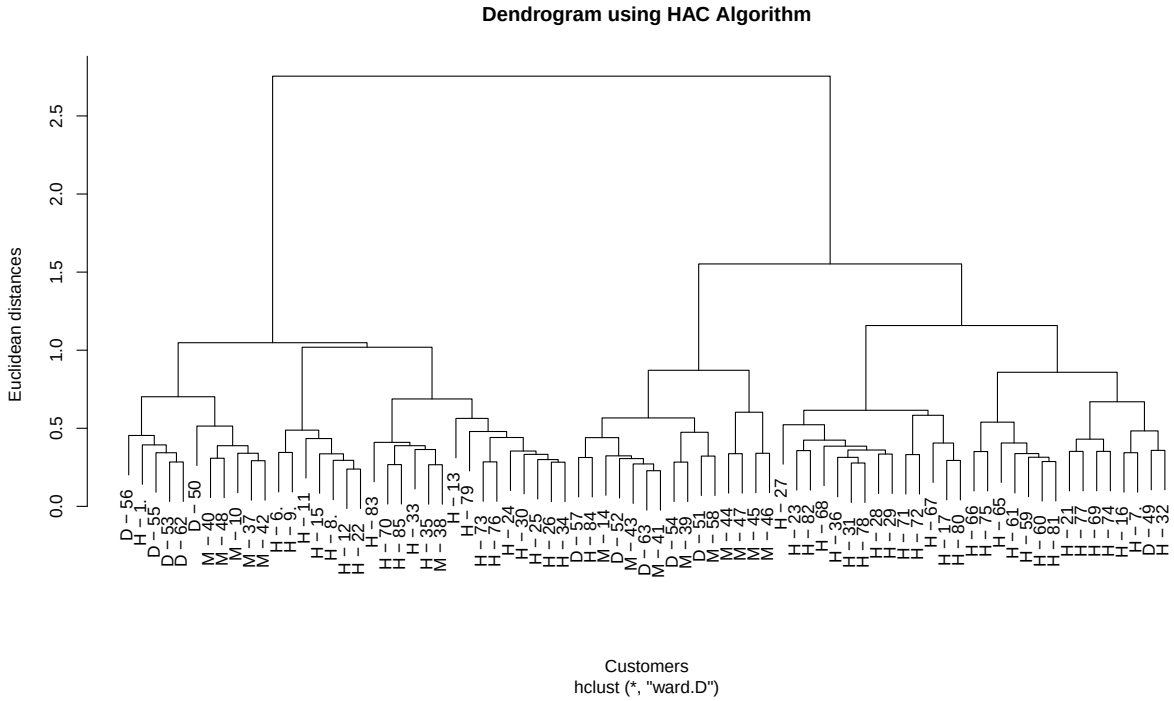


These two components explain 14.46 % of the point variability.

Hierarchical Clustering Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Using the dendrogram to find the optimal number of clusters

```
dendrogram = hclust(d = dist(subset2[1:(length(subset2)-1)], method = 'euclidean'), method = 'ward.D')
plot(dendrogram,
     main = paste('Dendrogram using HAC Algorithm'),
     xlab = 'Customers',
     ylab = 'Euclidean distances')
```



If we observe the dendogram carefully, we see that the disputed articles in each tail is associated with Hamilton

Fitting Hierarchical Clustering to the dataset

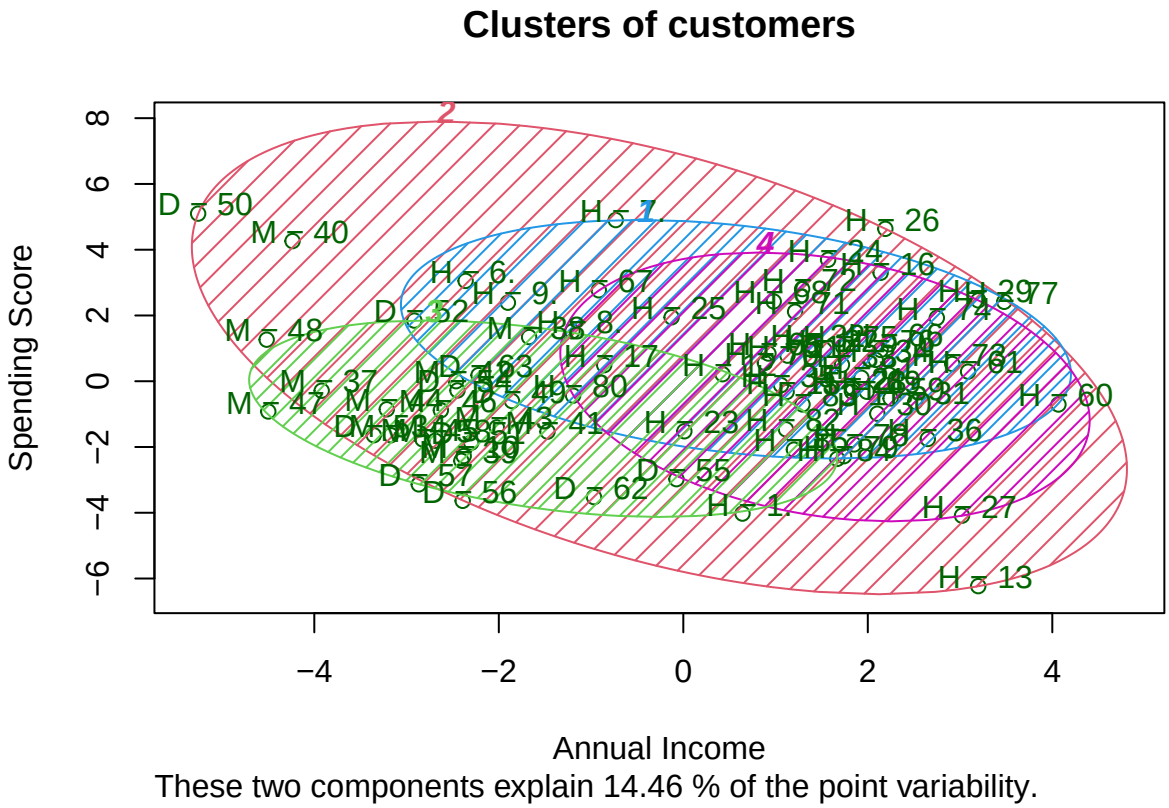
```
hc = hclust(d = dist(subset, method = 'euclidean'), method = 'ward.D')
```

```
## Warning in dist(subset, method = "euclidean"): NAs introduced by coercion
```

```
y_hc = cutree(hc, 4)
```

Visualising the clusters

```
library(cluster)
clusplot(subset,
         y_hc,
         lines = 0,
         shade = TRUE,
         color = TRUE,
         labels= 2,
         plotchar = FALSE,
         span = TRUE,
         main = paste('Clusters of customers'),
         xlab = 'Annual Income',
         ylab = 'Spending Score')
```



Conclusion: The disputed articles were authored by Madison as confirmed by K-Means and HAC.