Assignment 4

Dixitha Kasturi
dkasturi@syr.edu

Topic: Clustering Techniques

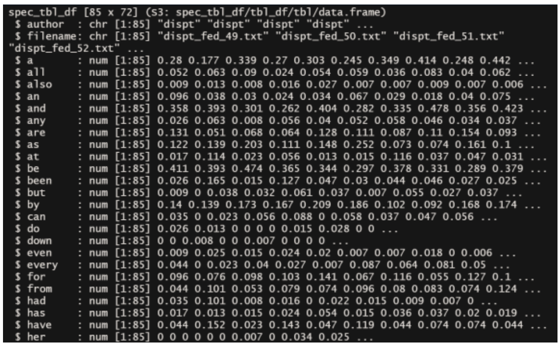
Dataset: Federalist Papers

Outcome: Cluster disputed documents correctly.

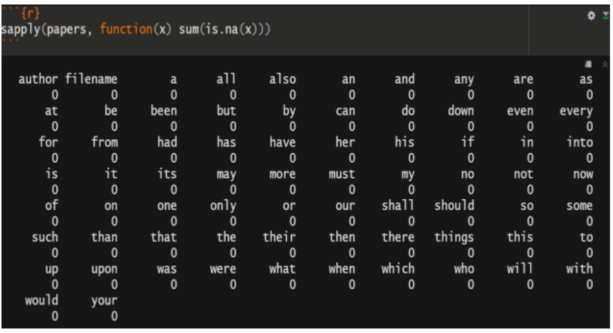
Exploratory Data Analysis:

85 papers with 72 attributes. Out of which the first two columns have the author name and the 2nd one has the actual file name of the paper. The other columns are counts of characters/words according to the name of the column.

Name of author and filename are both nominal attributes. This can be seen through the describe function.



```
spec_tbl_df [85 x 72] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ author  : chr [1:85] "dispt" "dispt" "dispt" "dispt" ...
 $ filename: chr [1:85] "dispt_fed_49.txt" "dispt_fed_50.txt" "dispt_fed_51.txt"
"dispt_fed_52.txt" ...
 $ a       : num [1:85] 0.28 0.177 0.339 0.27 0.303 0.245 0.349 0.414 0.248 0.442 ...
 $ all     : num [1:85] 0.052 0.063 0.09 0.024 0.054 0.059 0.036 0.083 0.04 0.062 ...
 $ also    : num [1:85] 0.009 0.013 0.008 0.016 0.027 0.007 0.007 0.009 0.007 0.006 ...
 $ an      : num [1:85] 0.096 0.038 0.03 0.024 0.034 0.067 0.029 0.018 0.04 0.075 ...
 $ and     : num [1:85] 0.358 0.393 0.301 0.262 0.404 0.282 0.335 0.478 0.356 0.423 ...
 $ any     : num [1:85] 0.026 0.063 0.008 0.056 0.04 0.052 0.058 0.046 0.034 0.037 ...
 $ are     : num [1:85] 0.131 0.051 0.068 0.064 0.128 0.111 0.087 0.11 0.154 0.093 ...
 $ as      : num [1:85] 0.122 0.139 0.203 0.111 0.148 0.252 0.073 0.074 0.161 0.1 ...
 $ at      : num [1:85] 0.017 0.114 0.023 0.056 0.013 0.015 0.116 0.037 0.047 0.031 ...
 $ be      : num [1:85] 0.411 0.393 0.474 0.365 0.344 0.297 0.378 0.331 0.289 0.379 ...
 $ been    : num [1:85] 0.026 0.165 0.015 0.127 0.047 0.03 0.044 0.046 0.027 0.025 ...
 $ but     : num [1:85] 0.009 0 0.038 0.032 0.061 0.037 0.007 0.055 0.027 0.037 ...
 $ by      : num [1:85] 0.14 0.139 0.173 0.167 0.209 0.186 0.102 0.092 0.168 0.174 ...
 $ can     : num [1:85] 0.035 0 0.023 0.056 0.088 0 0.058 0.037 0.047 0.056 ...
 $ do      : num [1:85] 0.026 0.013 0 0 0 0 0.015 0.028 0 0 ...
 $ down    : num [1:85] 0 0 0.008 0 0 0.007 0 0 0 0 ...
 $ even    : num [1:85] 0.009 0.025 0.015 0.024 0.02 0.007 0.007 0.018 0 0.006 ...
 $ every   : num [1:85] 0.044 0 0.023 0.04 0.027 0.007 0.087 0.064 0.081 0.05 ...
 $ for     : num [1:85] 0.096 0.076 0.098 0.103 0.141 0.067 0.116 0.055 0.127 0.1 ...
 $ from    : num [1:85] 0.044 0.101 0.053 0.079 0.074 0.096 0.08 0.083 0.074 0.124 ...
 $ had     : num [1:85] 0.035 0.101 0.008 0.016 0 0.022 0.015 0.009 0.007 0 ...
 $ has     : num [1:85] 0.017 0.013 0.015 0.024 0.054 0.015 0.036 0.037 0.02 0.019 ...
 $ have    : num [1:85] 0.044 0.152 0.023 0.143 0.047 0.119 0.044 0.074 0.074 0.044 ...
 $ her     : num [1:85] 0 0 0 0 0 0 0.007 0 0.034 0.025 ...
```



~~~
```{r}
sapply(papers, function(x) sum(is.na(x)))
```
~~~

```
  author filename        a      all     also       an      and      any      are       as
       0        0        0        0        0        0        0        0        0        0
      at       be     been      but       by      can       do     down     even    every
       0        0        0        0        0        0        0        0        0        0
     for     from      had      has     have      her      his       if       in     into
       0        0        0        0        0        0        0        0        0        0
      is       it      its      may     more     must       my       no      not      now
       0        0        0        0        0        0        0        0        0        0
      of       on      one     only       or      our    shall   should       so     some
       0        0        0        0        0        0        0        0        0        0
    such     than     that      the    their     then    there   things     this       to
       0        0        0        0        0        0        0        0        0        0
      up     upon      was     were     what     when    which      who     will     with
       0        0        0        0        0        0        0        0        0        0
   would     your
       0        0
```

We see that the data is clean and has no missing values. There are 3 different authors and few combined by Hamilton and Madison and 85 text files. Out of these 85 we have 11 disputed records. We are supposed to find which cluster these files would belong to.

51 essays written by Hamilton

15 by Madison

3 by Hamilton and Madison

5 by Jay

Clustering Analysis

After doing analysis including Jay's articles and excluding Jay's articles, I divided the report into those two sections:

Section1: Including Jay's articles

Section 2: Excluding Jay's articles

Section 1 : Including Jay's articles

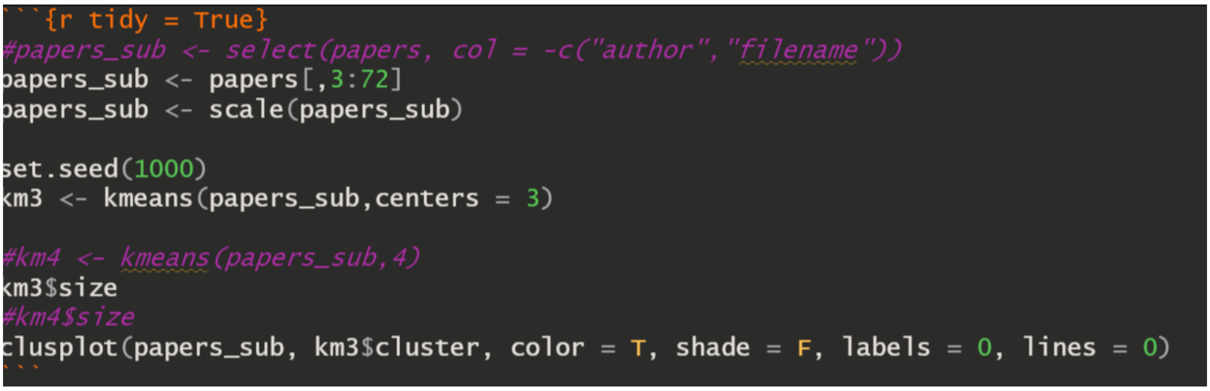
We will do 2 types of clustering. Kmeans and Hierarchical clustering

Method 1 : Kmeans

We want 3 clusters as we have 3 different authors.

We will use the columns which have counts of words to understand the document distribution, as authors have a style of writing, we subset the data

We then scale our data so that clustering doesn't depend only on one attribute.

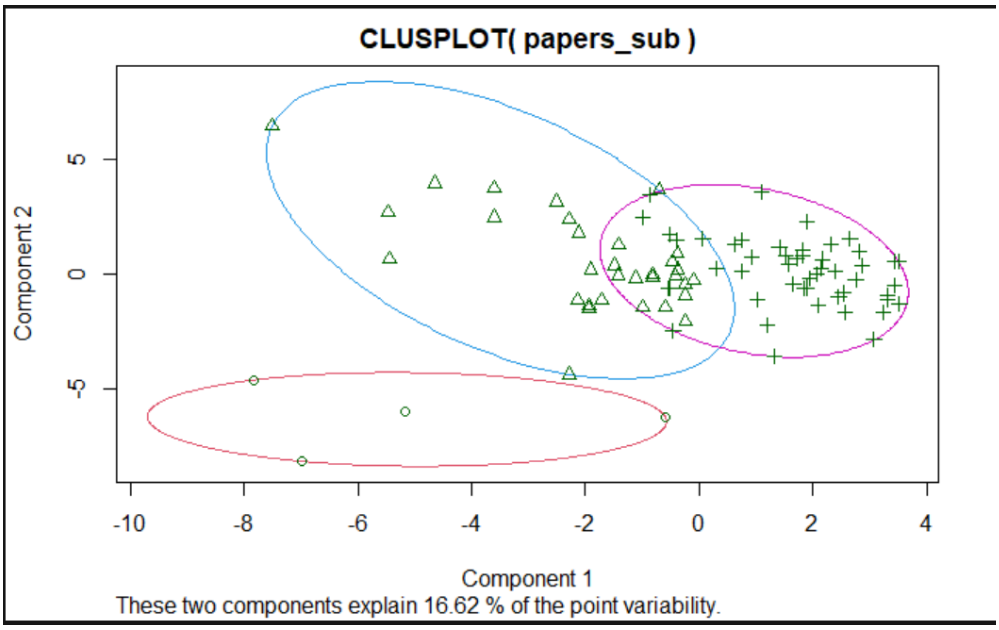


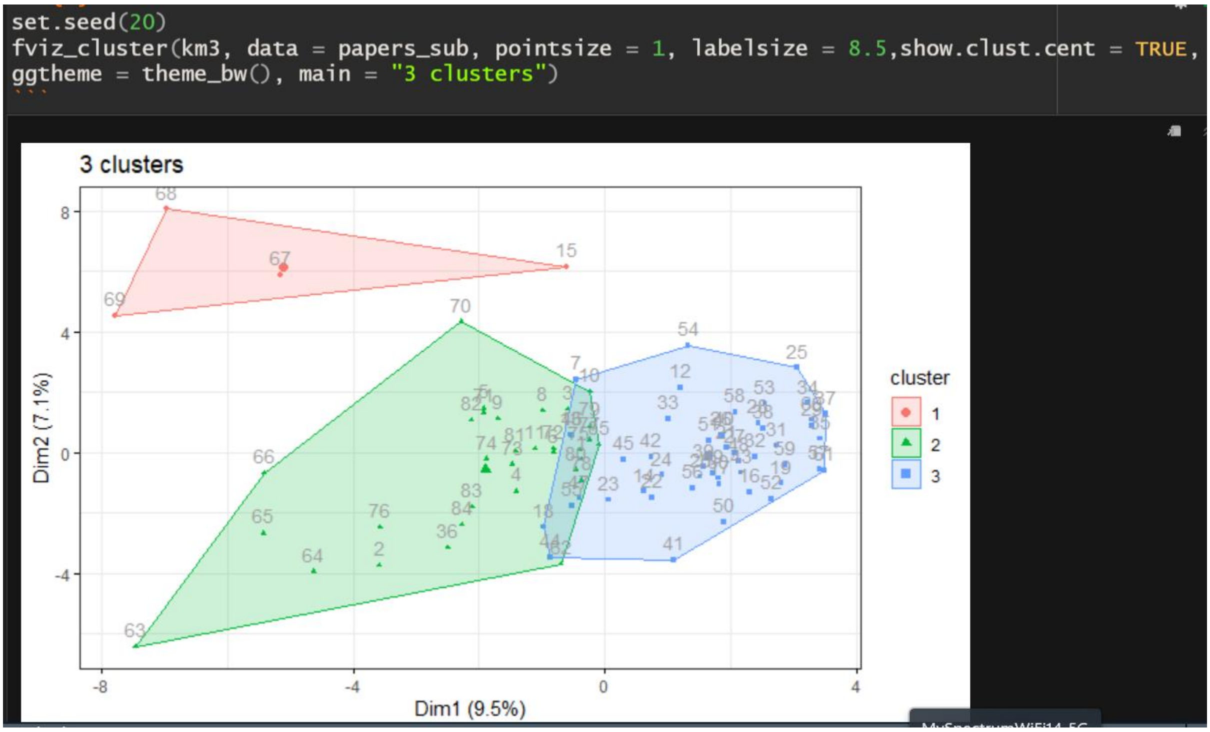
~~~
```{r tidy = True}
#papers_sub <- select(papers, col = -c("author","filename"))
papers_sub <- papers[,3:72]
papers_sub <- scale(papers_sub)

set.seed(1000)
km3 <- kmeans(papers_sub,centers = 3)

#km4 <- kmeans(papers_sub,4)
km3$size
#km4$size
clusplot(papers_sub, km3$cluster, color = T, shade = F, labels = 0, lines = 0)
```
~~~

We see that the sizes of the clusters don't match to what we expected, even if we consider HM as a different cluster. We plot the clusters to see how clusters are exactly formed

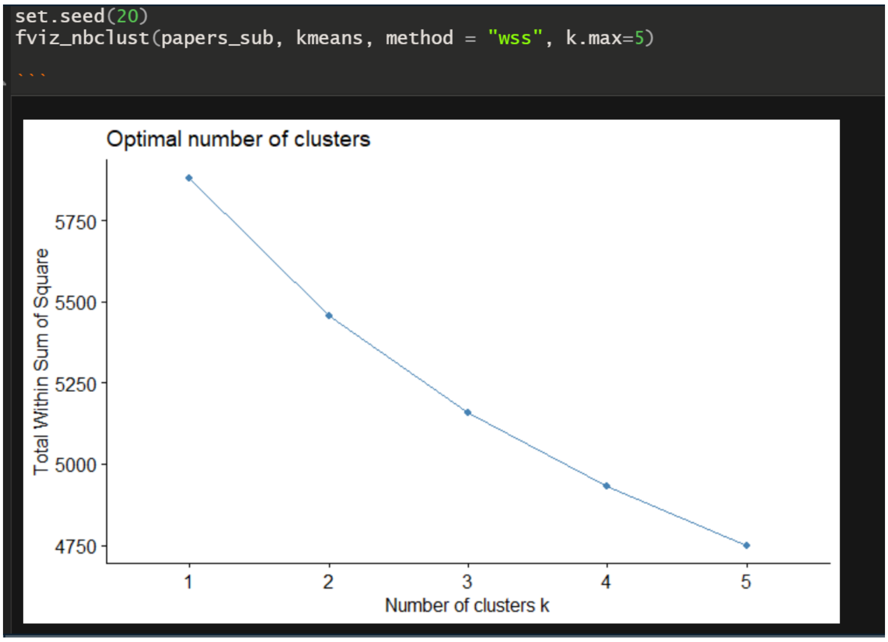




```
set.seed(20)
fviz_cluster(km3, data = papers_sub, pointsize = 1, labelsize = 8.5,show.clust.cent = TRUE,
ggtheme = theme_bw(), main = "3 clusters")
```

We clearly see how the clusters overlap and the documents on manual verification are not clustered the way they are supposed to. Even if we change the number of clusters, it still didn't give good results.

We choose the optimum number of clusters that need to be formed using Elbow method. It is still not very clear, as to how many clusters should be taken. Hence, we stick with 3.



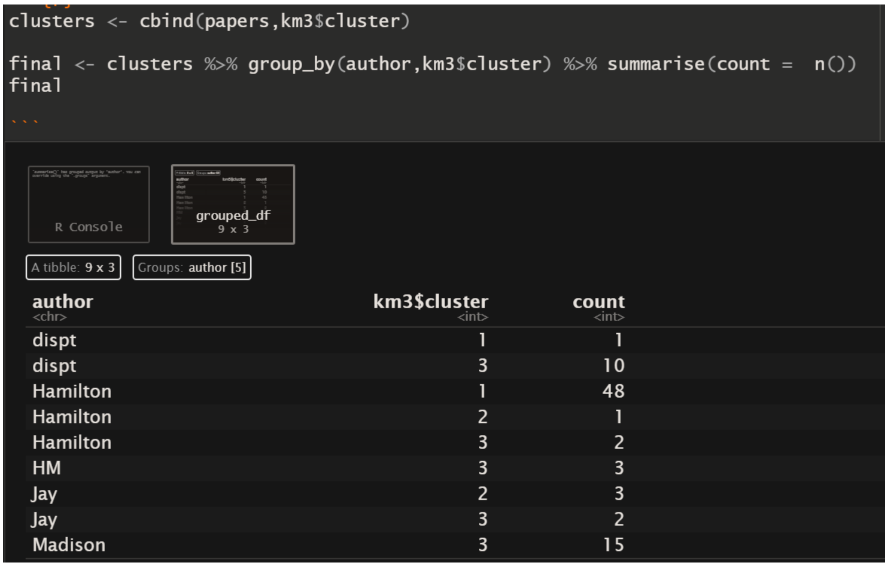
```
set.seed(20)
fviz_nbclust(papers_sub, kmeans, method = "wss", k.max=5)

```
```

Optimal number of clusters
Total Within Sum of Square
5750
5500
5250
5000
4750
1
2
3
4
5
Number of clusters k

On checking the cluster results, we see that there is a lot of overlapping happening. The clusters that are formed are not very clear and there is a lot of overlapping. Though we see that Madison's papers are all grouped together in a single cluster3, but this cluster is overlapped with everything.

We see that 1 disputed article is in cluster 1 which has only Hamilton works. We can be sure that this paper was written by Hamilton. It is unclear by who the other 10 were written. We see if hierarchical clustering would clear out the ambiguity about the disputed articles.



```
clusters <- cbind(papers,km3$cluster)

final <- clusters %>% group_by(author,km3$cluster) %>% summarise(count =  n())
final

```
```

A tibble: 9 x 3 | Groups: author [5]

author <chr>	km3$cluster <int>	count <int>
dispt	1	1
dispt	3	10
Hamilton	1	48
Hamilton	2	1
Hamilton	3	2
HM	3	3
Jay	2	3
Jay	3	2
Madison	3	15

Method 2 : Hierarchical Clustering

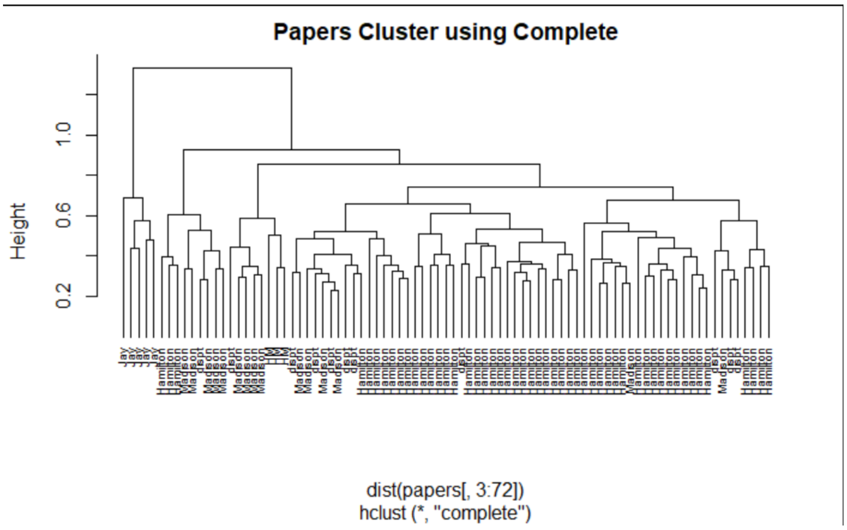
a) using complete method

b) using average method

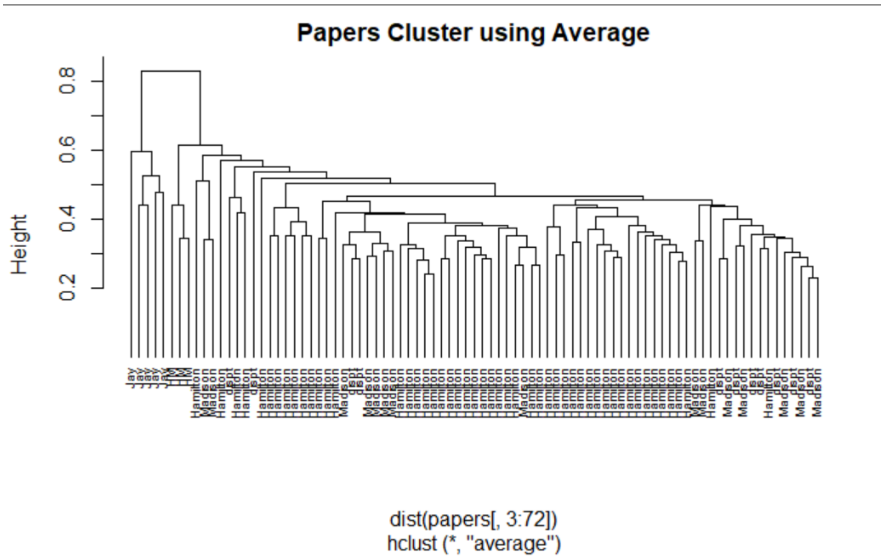
c) using single method

When we use complete and average methods in hierarchical clustering, we see that Jay was grouped together but in single method the results were a little like what we saw from kmeans. While it is unclear, going by the majority, we see that HAC performed better in clustering the files. Jay were grouped. In complete and average method - there was overlap between hamilton, madison and disputed articles. So, we can be sure here that the articles were written by either of them. The same went for single method, but the only difference was that jay articles were in different clusters.

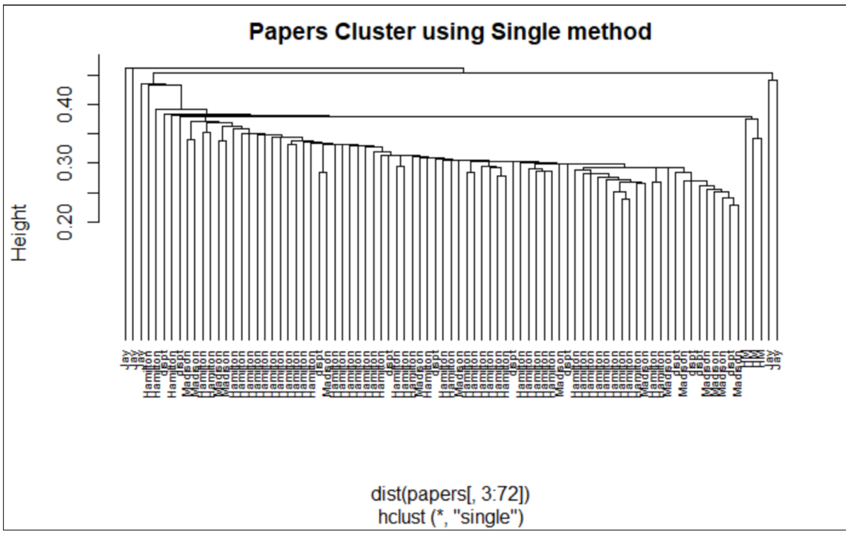
a) Complete method:



b) Average method:



c) Single method:

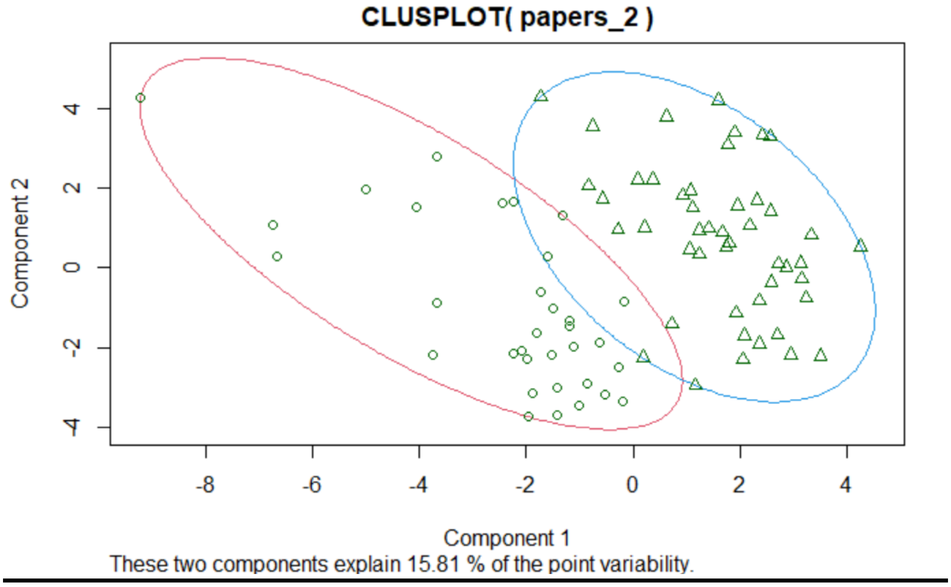


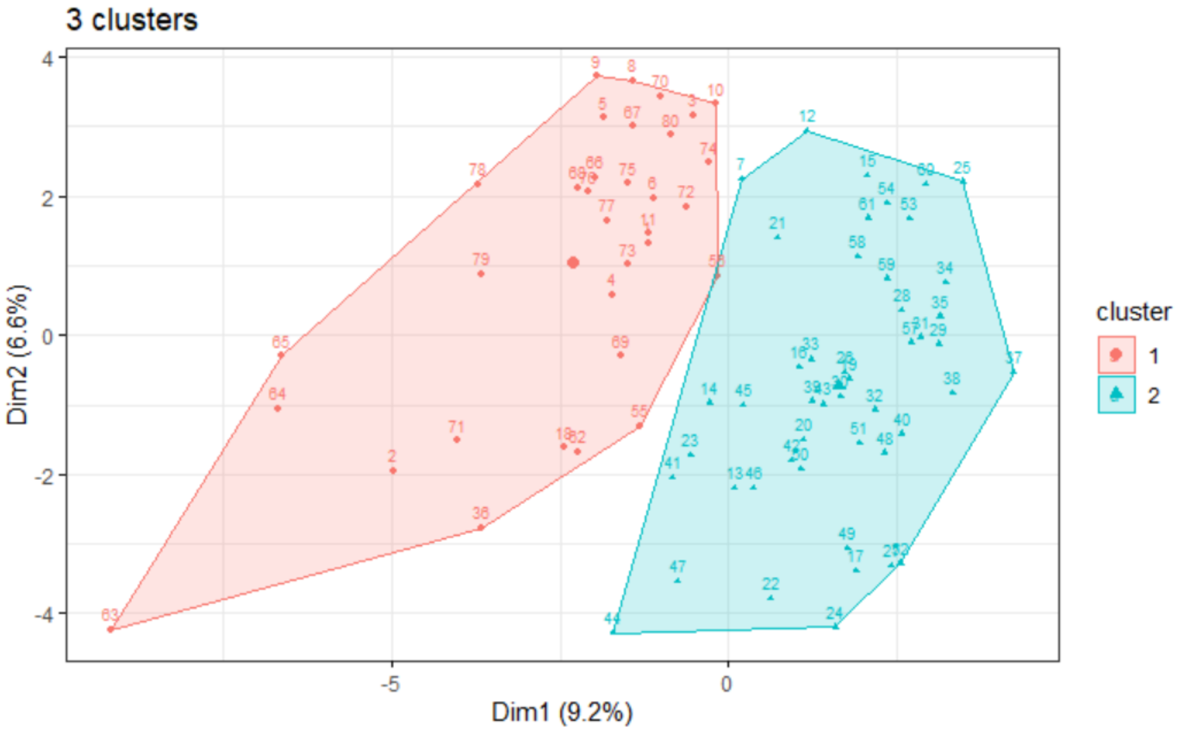
Section 2 : Excluding Jay's articles

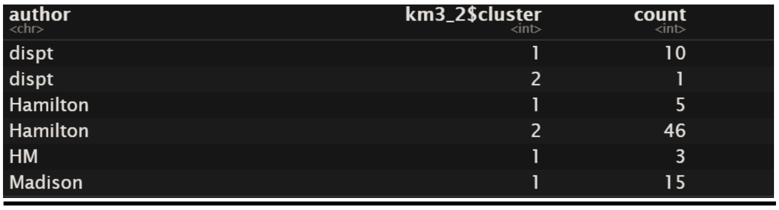
We will do 2 types of clustering. Kmeans and Hierarchical clustering

Method 1 : Kmeans

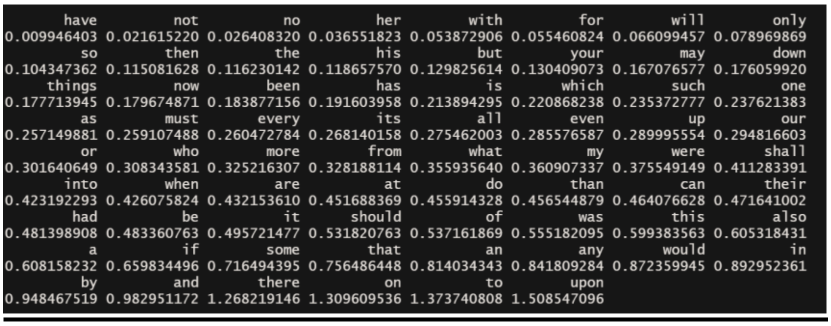
The blue cluster is representing Hamilton whereas the red one represents Madison. the disputed files here are clustered closer to Madison's center. The word that had highest weightage in clustering is "upon". There is 1disputed article that is falling under Hamilton but the majority are under madison's cluster. The joint authorship papers(64,65,63) are located in the red cluster far from the center of the red cluster.







author <chr>	km3_2$cluster <int>	count <int>
dispt	1	10
dispt	2	1
Hamilton	1	5
Hamilton	2	46
HM	1	3
Madison	1	15



```
       have         not          no         her        with         for        will        only 
0.009946403 0.021615220 0.026408320 0.036551823 0.053872906 0.055460824 0.066099457 0.078969869 
         so        then         the         his         but        your         may        down 
0.104347362 0.115081628 0.116230142 0.118657570 0.129825614 0.130409073 0.167076577 0.176059920 
     things         now        been         has          is       which        such         one 
0.177713945 0.179674871 0.183877156 0.191603958 0.213894295 0.220868238 0.235372777 0.237621383 
         as        must       every         its         all        even          up         our 
0.257149881 0.259107488 0.260472784 0.268140158 0.275462003 0.285576587 0.289995554 0.294816603 
         or         who        more        from        what          my        were       shall 
0.301640649 0.308343581 0.325216307 0.328188114 0.355935640 0.360907337 0.375549149 0.411283391 
       into        when         are          at          do        than         can       their 
0.423192293 0.426075824 0.432153610 0.451688369 0.455914328 0.456544879 0.464076628 0.471641002 
        had          be          it      should          of         was        this        also 
0.481398908 0.483360763 0.495721477 0.531820763 0.537161869 0.555182095 0.599383563 0.605318431 
          a          if        some        that          an         any       would          in 
0.608158232 0.659834496 0.716494395 0.756486448 0.814034343 0.841809284 0.872359945 0.892952361 
         by         and       there          on          to        upon 
0.948467519 0.982951172 1.268219146 1.309609536 1.373740808 1.508547096 
```

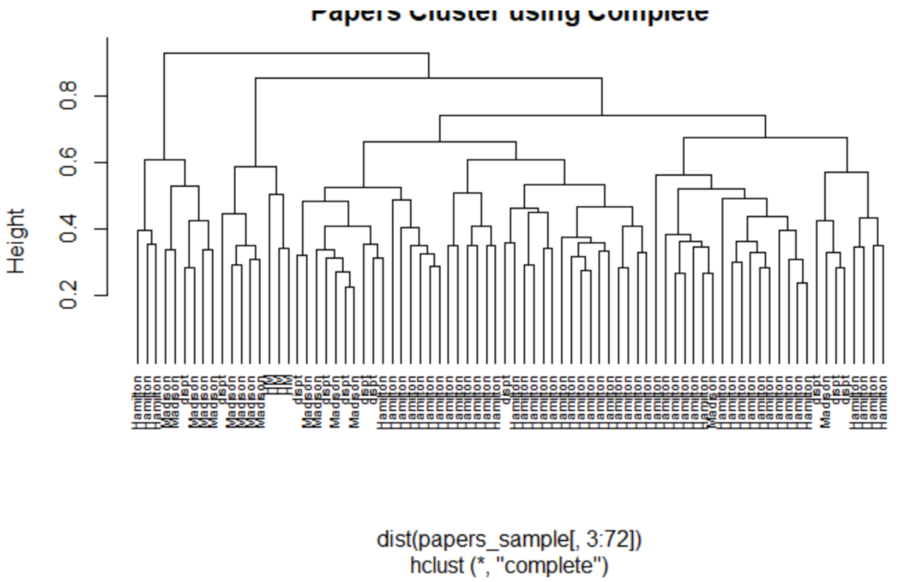
Method 2 : Hierarchical Clustering

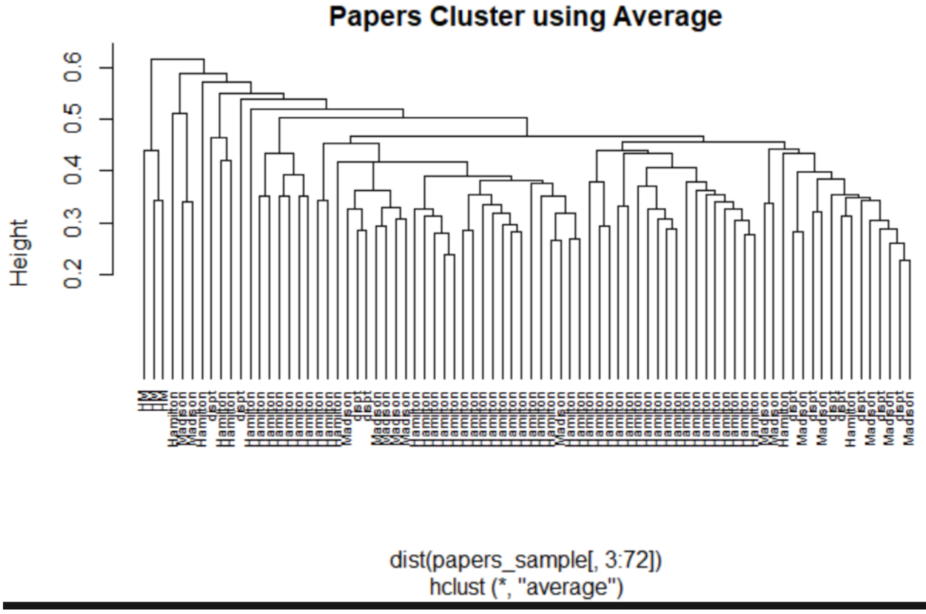
a) using complete method

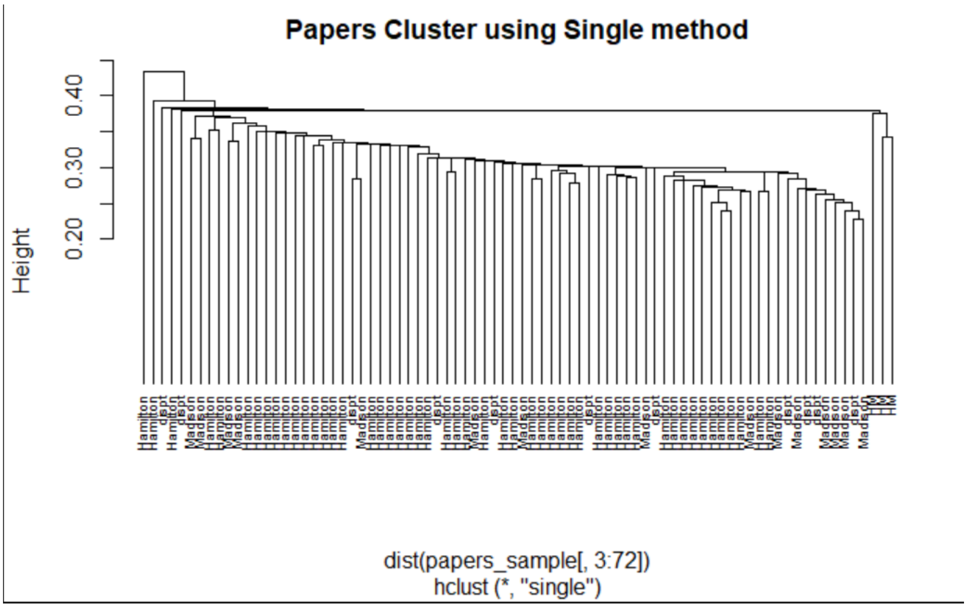
b) using average method

c) using single method

On looking at all 3 methods, the disputed articles are leaning more towards Madison clustered branches.


Papers Cluster using Complete
Height
0.2
0.4
0.6
0.8
dist(papers_sample[, 3:72])
hclust (*, "complete")


Papers Cluster using Average
Height
0.2
0.3
0.4
0.5
0.6
dist(papers_sample[, 3:72])
hclust (*, "average")



Conclusion:

According to this report, the disputed papers belong to Madison from both HAC(all 3 methods) & Kmeans analysis.

We see that out of all the attributes the word “upon” is the most useful for clustering and differentiating the papers. The centroid values for the ‘upon’ dimension are the farthest.

If we included the papers from Jay, there was a lot of distortion and the results were not accurate. When these papers were excluded exactly according to the claims of the report, we see that the disputed files leaned more towards madison. So we can say that the articles belonged to madison.