Homework 4 - Olivieri

Code ▾

Hide

```
#### Replace working directory as necessary
setwd(getwd())
getwd()
```

```
[1] "/Users/aoliv01/Desktop/GradSchool/2018-2/DataMining/Homework/HW4"
```

Hide

```
### Read in the file
raw <- read.csv('fedPapers85.csv')
```

Hide

```
cat('Number of NA\'s: ', sum(is.na(raw)))
```

```
Number of NA's:  0
```

Hide

```
### Check the dataframe for variable type
str(raw)
```

Hide

```
### Getting a look at the column names
### The words are after column 1 and 2
colnames(raw)
```

Hide

```
### Removing the filename and author column
words <- raw[,-1:-2]
```

Hide

```
### Setting k means to 4 centroids
### These centroids will represent the 3 authors and the mixed author of HM
## Set the seed for reproducibility
set.seed(1234)
m_k <- kmeans(words, 5, iter.max = 5000)
```

Hide

```
### There's going to be a lot of noise in here
### The word values aren't a nominal (binary) value
### Also the words are common words: 'at', 'are', 'shall'
### The results are going to be mixed-up
m_k$centers
```

```
     a   all   also    an  and   any   are   as    at   be  been   but   by   can     do
   down
1 0.28 0.057 0.0076 0.073 0.34 0.044 0.074 0.13 0.038 0.33 0.059 0.032 0.12 0.039 0.0058
0.0024
2 0.16 0.036 0.0198 0.025 0.72 0.038 0.085 0.16 0.036 0.28 0.027 0.049 0.14 0.033 0.0082
0.0000
3 0.32 0.051 0.0062 0.069 0.39 0.039 0.079 0.11 0.051 0.28 0.065 0.031 0.14 0.033 0.0065
0.0009
    even every  for.  from   had   has  have    her   his   if.  in.  into    is   it
its   may
1 0.0109 0.025 0.091 0.074 0.015 0.039 0.093 0.0022 0.022 0.026 0.33 0.020 0.167 0.17 0.
053 0.067
2 0.0076 0.006 0.096 0.091 0.016 0.029 0.087 0.0148 0.009 0.053 0.27 0.045 0.094 0.20 0.
033 0.057
3 0.0123 0.025 0.096 0.084 0.027 0.052 0.097 0.0129 0.037 0.025 0.31 0.026 0.153 0.14 0.
045 0.057
   more  must     my    no   not    now   of    on   one  only    or   our shall should
   so  some
1 0.041 0.032 0.0024 0.036 0.092 0.0055 0.94 0.069 0.036 0.025 0.096 0.012 0.021  0.029
0.028 0.016
2 0.087 0.021 0.0018 0.015 0.108 0.0066 0.64 0.075 0.081 0.043 0.161 0.066 0.017  0.041
0.045 0.021
3 0.045 0.035 0.0043 0.031 0.091 0.0064 0.91 0.069 0.040 0.019 0.090 0.028 0.017  0.022
0.030 0.023
   such  than that  the their   then there things  this   to     up   upon   was  were
what  when
1 0.028 0.040 0.22 1.44 0.074 0.0057 0.028 0.0028 0.090 0.56 0.0012 0.0313 0.023 0.017
0.014 0.011
2 0.051 0.063 0.24 0.85 0.142 0.0080 0.014 0.0014 0.053 0.48 0.0000 0.0018 0.025 0.029
0.018 0.021
3 0.028 0.046 0.20 1.19 0.089 0.0062 0.026 0.0027 0.088 0.52 0.0061 0.0306 0.028 0.022
0.012 0.012
  which   who  will  with would    your
1 0.164 0.028 0.109 0.077  0.10 0.00087
2 0.099 0.052 0.126 0.095  0.13 0.00640
3 0.159 0.034 0.085 0.080  0.10 0.00259
```

Hide

```
### We can take the original data frame with the author attribute
### and attach to a new data frame with the cluster
word_cluster <- data.frame(raw, m_k$cluster)
head(word_cluster)
```

	author <fctr>	filename <fctr>	a <dbl>	all <dbl>	also <dbl>	an <dbl>	and <dbl>	any <dbl>	are <dbl>
1	dispt	dispt_fed_49.txt	0.28	0.052	0.009	0.096	0.36	0.026	0.131
2	dispt	dispt_fed_50.txt	0.18	0.063	0.013	0.038	0.39	0.063	0.051
3	dispt	dispt_fed_51.txt	0.34	0.090	0.008	0.030	0.30	0.008	0.068
4	dispt	dispt_fed_52.txt	0.27	0.024	0.016	0.024	0.26	0.056	0.064
5	dispt	dispt_fed_53.txt	0.30	0.054	0.027	0.034	0.40	0.040	0.128
6	dispt	dispt_fed_54.txt	0.24	0.059	0.007	0.067	0.28	0.052	0.111

6 rows | 1-10 of 73 columns

Hide

```
tail(word_cluster)
```

	author <fctr>	filename <fctr>	a <dbl>	all <dbl>	also <dbl>	an <dbl>	and <dbl>	any <dbl>	are <dbl>
80	Madison	Madison_fed_44.txt	0.21	0.101	0.010	0.051	0.44	0.076	0.066
81	Madison	Madison_fed_45.txt	0.14	0.054	0.014	0.048	0.42	0.027	0.048
82	Madison	Madison_fed_46.txt	0.21	0.028	0.006	0.050	0.39	0.033	0.073
83	Madison	Madison_fed_47.txt	0.18	0.052	0.047	0.047	0.44	0.026	0.135
84	Madison	Madison_fed_48.txt	0.24	0.091	0.008	0.084	0.37	0.008	0.046
85	Madison	Madison_fed_58.txt	0.35	0.097	0.007	0.056	0.31	0.035	0.049

6 rows | 1-10 of 73 columns

Hide

```
table(word_cluster$author, m_k$cluster)
```

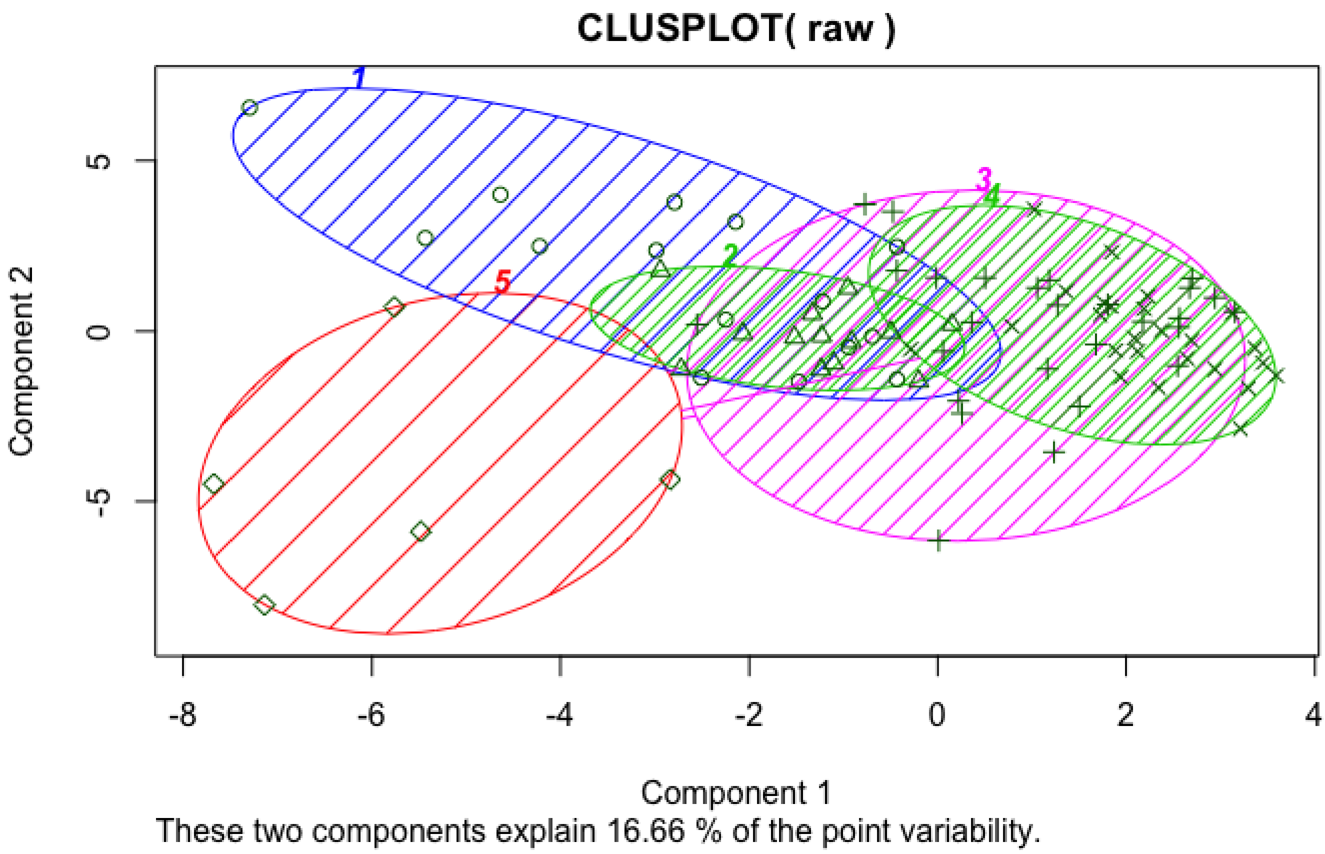
```
           1  2  3  4  5
dispt      4  5  2  0  0
Hamilton   2  0 24 25  0
HM         3  0  0  0  0
Jay        0  0  0  0  5
Madison    6  8  1  0  0
```

Hide

```
library(cluster)
```

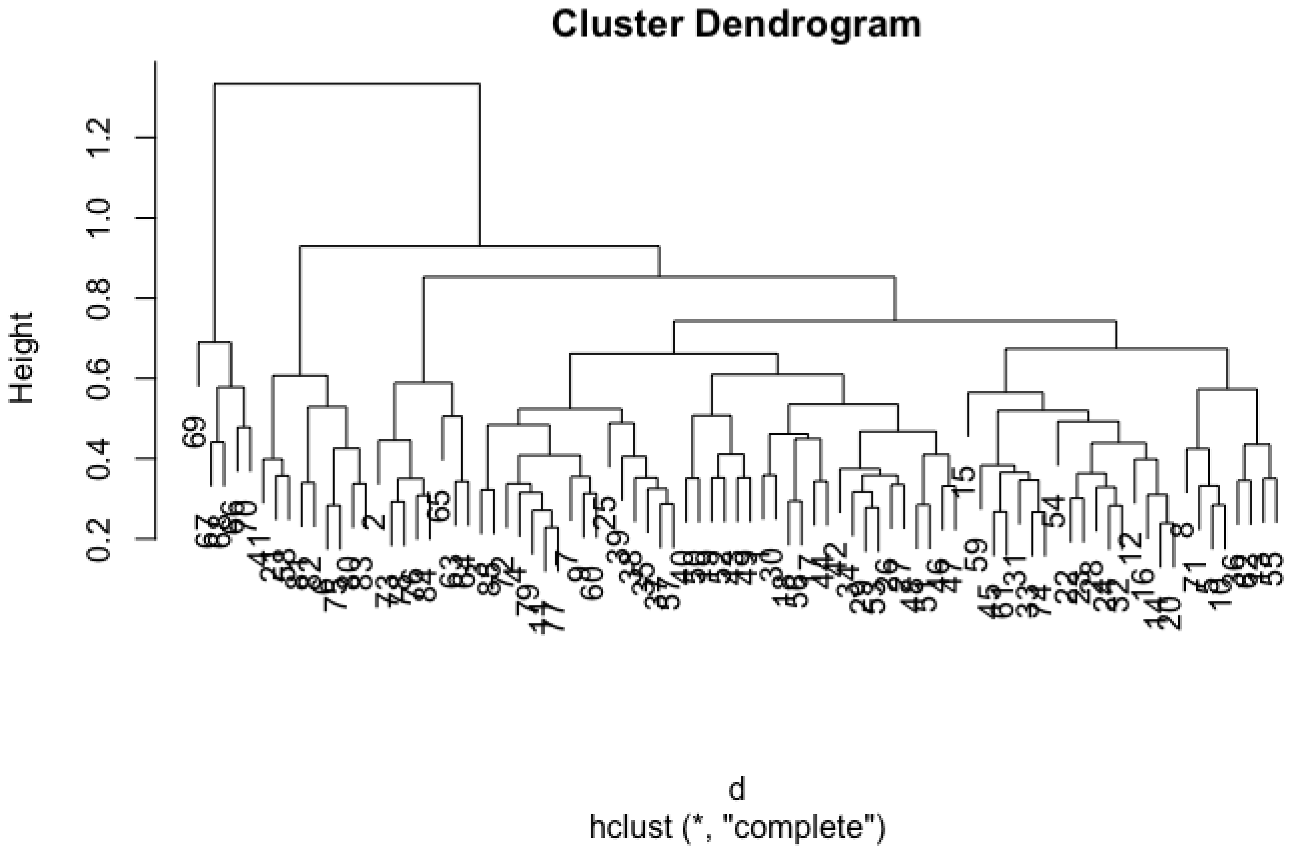
Hide

```
### Looking at a cluster plot shows Jay down on his own
### The other clusters representing Hamilton, Madison, HM, and disputed
### there's a lot of overlay
## Let's say cluster 1 is Madison and cluster 4 is Hamilton
## Cluster 3 is both of them, 'HM'
## Which makes cluster 2, our disputed author
## Cluster 2 seems to be wholly engulfed by Madison, sharing some values with 'HM', and
 Hamilton
clusplot(raw, m_k$cluster, color = T, shade = T, labels = 5, plotchar = T)
```



Hide

```
d = dist(as.matrix(words))
hc = hclust(d)
plot(hc)
```



Hide

```
cluster_cut <- cutree(hc, 5)
table(cluster_cut, type = raw$author)
```

```
           type
cluster_cut dispt Hamilton HM Jay Madison
          1     6       28  0   0       4
          2     1        0  3   0       4
          3     3       20  0   0       2
          4     1        3  0   0       5
          5     0        0  0   5       0
```

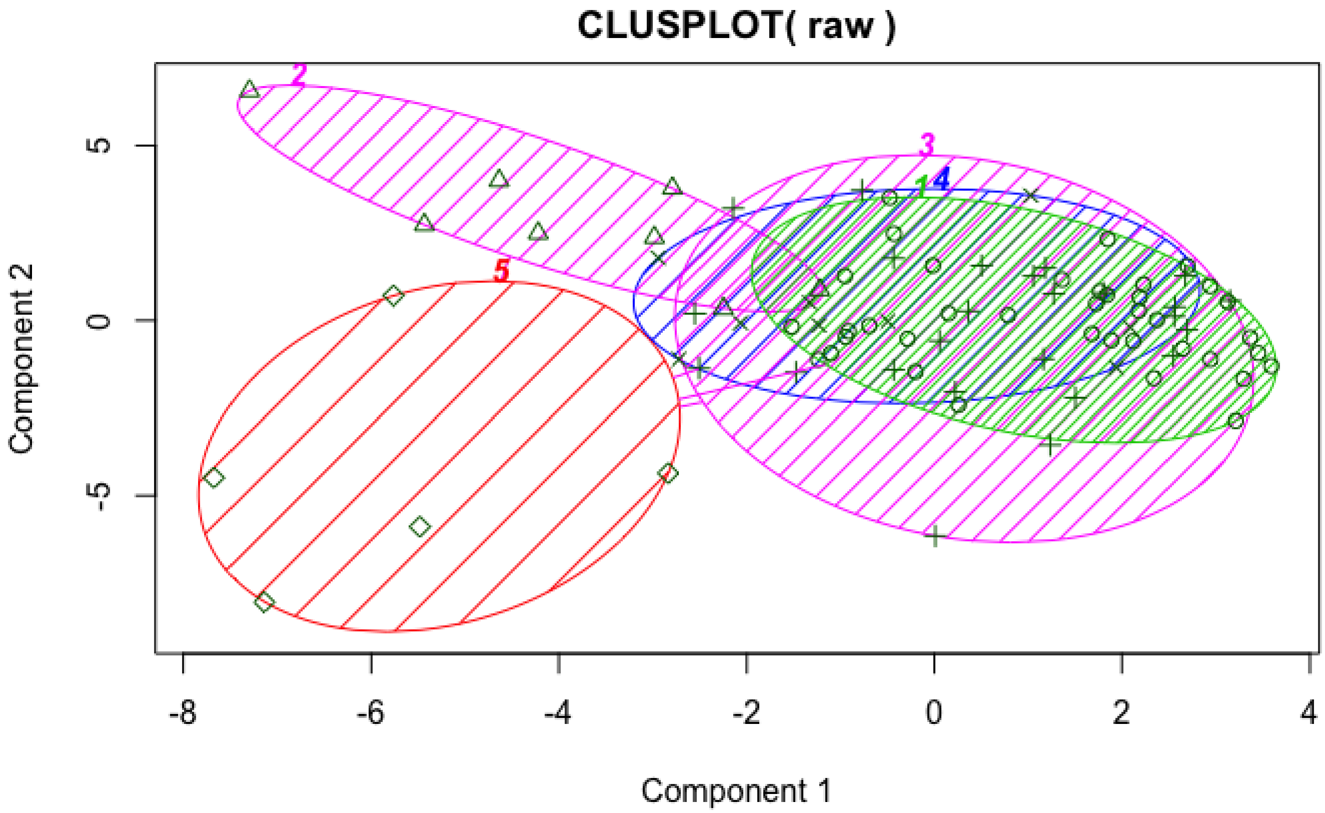
Hide

```
table(cluster = m_k$cluster, type = raw$author)
```

```
       type
cluster dispt Hamilton HM Jay Madison
      1     4        2  3   0       6
      2     5        0  0   0       8
      3     2       24  0   0       1
      4     0       25  0   0       0
      5     0        0  0   5       0
```

Hide

```
clusplot(raw, cluster_cut, color = T, shade = T, labels = 5, plotchar = T)
```



These two components explain 16.66 % of the point variability.

The cluster analysis reveals milky results. The commonality of the words and the similar writing styles of Hamilton and Madison made splitting the papers into distinct separate authors challenging. Only by small variances are Hamilton and Madison split away from each other. It’s not with great confidence, but I’m calling the disputed Federalist Papers to have been written primarily by Madison – it also wouldn’t be surprising that Hamilton wrote some or at least collaborated / edited the works.