k-means clustering and HAC

Load libraries and Set Knit

Load data

Today, we'll be working on the federalist papers data set. Papers written by Madison and Hamilton, and some papers where the author is unknown (assumed to be Madison or Hamilton). The text is "vectorized" – word counts are encoded as frequencies.

```
FederalistPapers <- read.csv("fedPapers85_fromClass.csv", na.strings = c(""))

# Create backup of FederalistPapers in case it's needed
FederalistPapers_Orig <- FederalistPapers

# Take a look at the data
View(FederalistPapers)

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

```
[1] 0
```

```
#-------------------------------------------------------------------------------------------------------------
```

K means

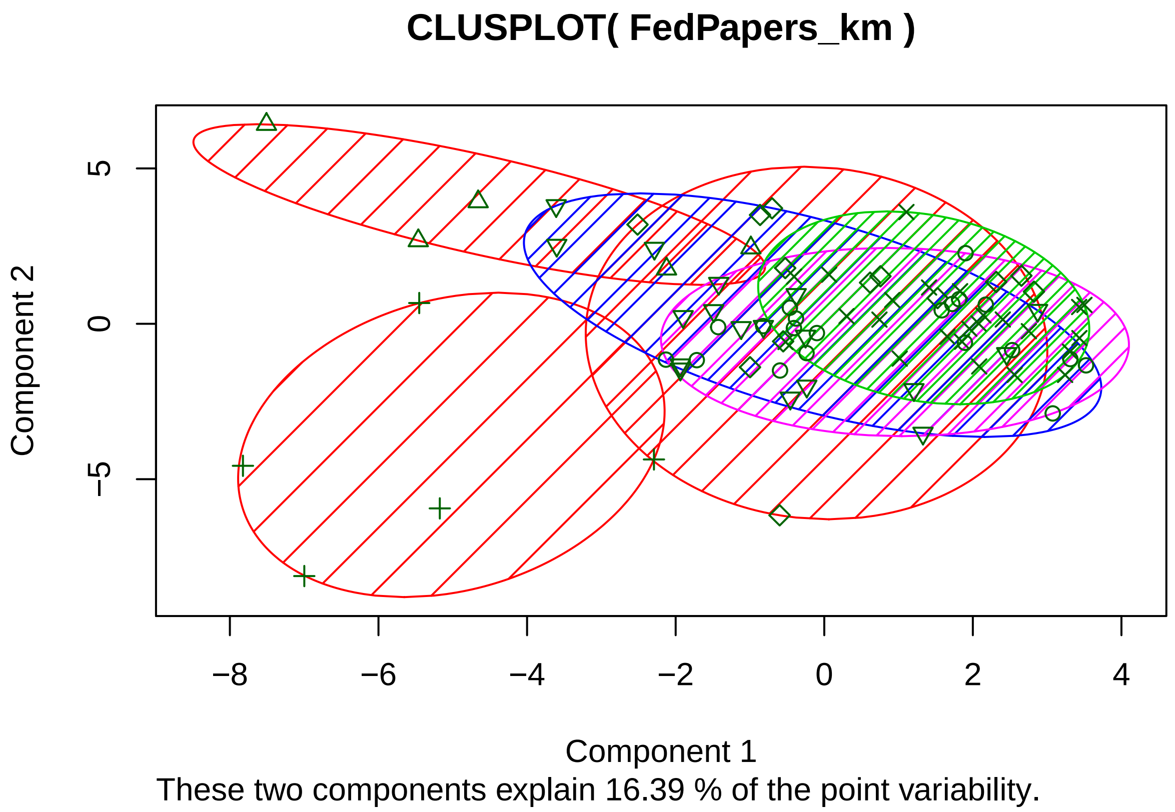
Once the data has been vectorized we can now apply analysis techniques such as clustering. First we will remove the labels and determine the "optimal" numbers of clusters for the clustering algorithm. Using total withinness we can estimate a "good" number of clusters.

```
# Remove author names from dataset for clustering purposes
FedPapers_km <-FederalistPapers[,2:72]

# Reduce the dimensionality ... focus on signal and not noise :)
#FedPapers_km <- select(FedPapers_km, filename, upon, all, may, also, even, from, shall, only)

# Make the file names the row names. Need a dataframe of numerical values for k-means
rownames(FedPapers_km) <- FedPapers_km[,1]
FedPapers_km[,1] <- NULL

#View(FedPapers_km)

# Determine "Optimal" number of clusters
# ANTIQUATED fviz_nbclust(FederalistPapers, FUN=hcut,method = "wss")
#fviz_nbclust(FederalistPapers, FUN =hcut, method = "silhouette")

# Set seed for fixed random seed
set.seed(20)

# run k-means
Clusters <- kmeans(FedPapers_km, 6)
FedPapers_km$Clusters <- as.factor(Clusters$cluster)
```

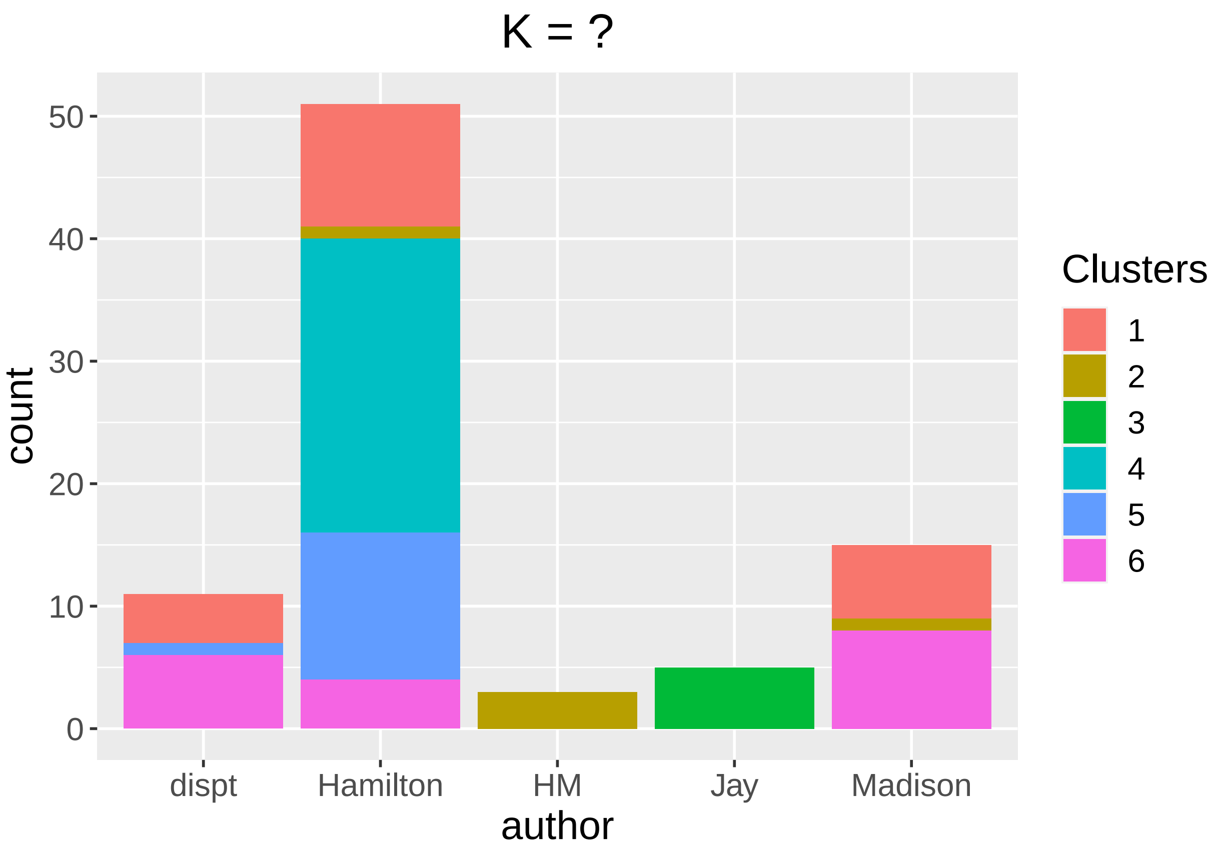
```r
str(Clusters)
Clusters$centers
```

Next we will add the clustering results back to the dataframe and display the findings. We can attempt to identify if the clustering results intuitively group papers written by the same author into the same clusters.

```r
# Add clusters to dataframe original dataframe with author name
FedPapers_km2 <- FederalistPapers
FedPapers_km2$Clusters <- as.factor(Clusters$cluster)

# Plot results
clusplot(FedPapers_km, FedPapers_km$Clusters, color=TRUE, shade=TRUE, labels=0, lines=0)
```



```r
ggplot(data=FedPapers_km2, aes(x=author, fill=Clusters))+
  geom_bar(stat="count") +
  labs(title = "K = ?") +
  theme(plot.title = element_text(hjust=0.5), text=element_text(size=15))
```



```
#------------------------------------------------------------------------------------------------------
```

Are these results good? Should we choose a different number of clusters?? Try to tweak / tune the model to improve results. What do you find??

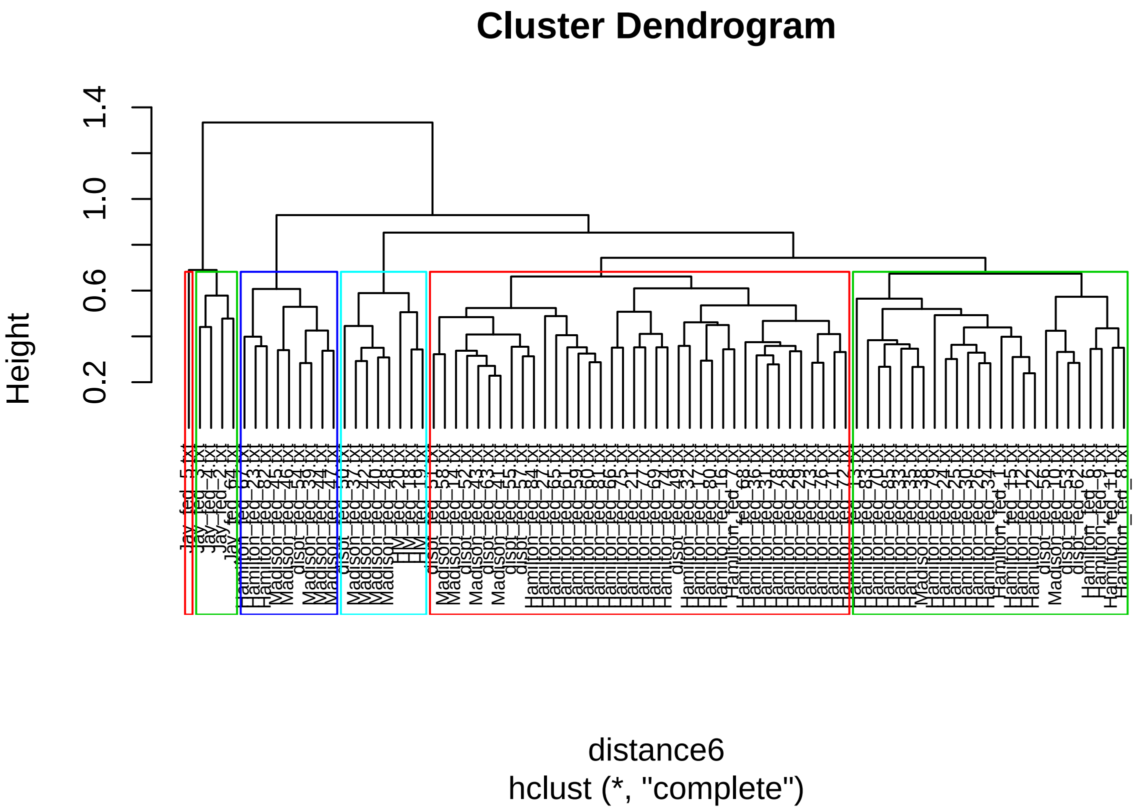
Hierachical Clustering Algorithms (HAC)

Next we will apply hierarchical clustering methods using various distance metrics.

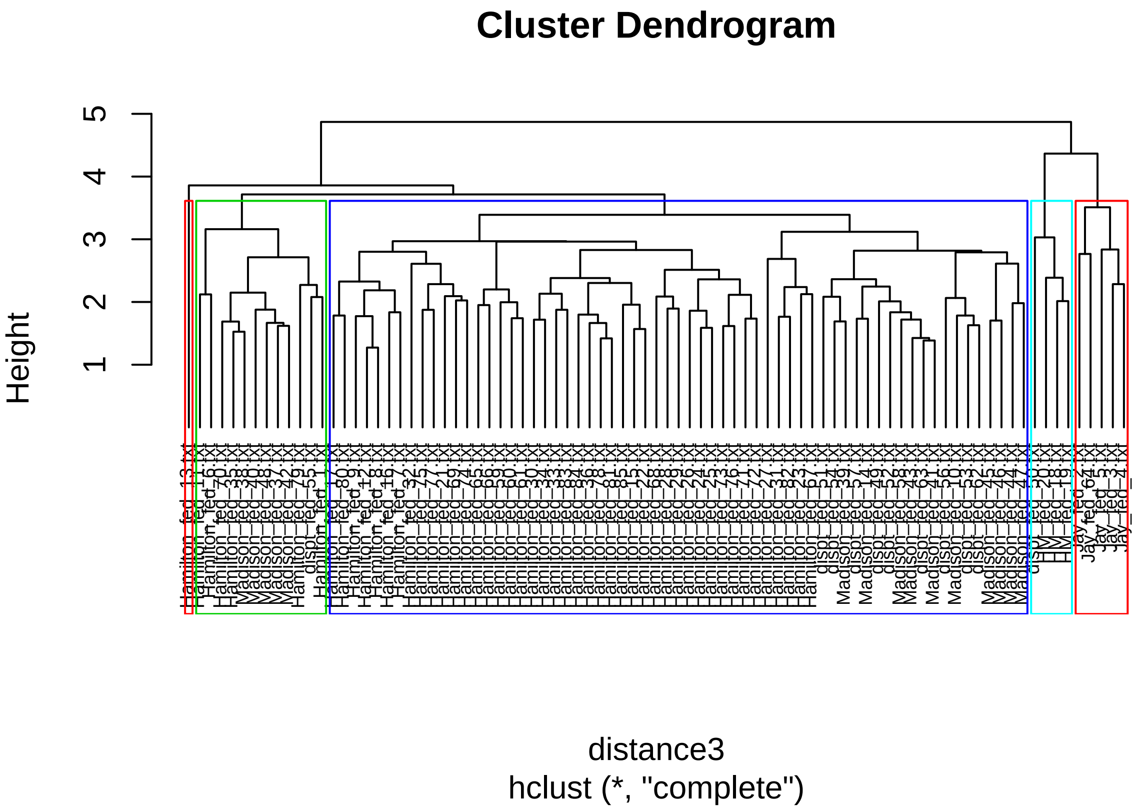
```
# Remove author names from dataset
FedPapers_HAC <- FederalistPapers[,c(2:72)]

# Make the file names the row names. Need a dataframe of numerical values for HAC
rownames(FedPapers_HAC) <- FedPapers_HAC[,1]
FedPapers_HAC[,1] <- NULL

View(FedPapers_HAC)

# Calculate distance in a variety of ways
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")
```

Below we display the results of HAC. The boxes indicate clusters.

```
HAC <- hclust(distance6, method="complete")
plot(HAC, cex=0.6, hang=-1)
rect.hclust(HAC, k =6, border=2:5)
```



```
HAC2 <- hclust(distance3, method="complete")
plot(HAC2, cex=0.6, hang=-1)
rect.hclust(HAC2, k =5, border=2:5)
```


Cluster Dendrogram
Height
1
2
3
4
5
Hamilton_fed_13.txt
Hamilton_fed_11.txt
Hamilton_fed_16.txt
Hamilton_fed_70.txt
Hamilton_fed_35.txt
Madison_fed_38.txt
Madison_fed_40.txt
Madison_fed_48.txt
Madison_fed_47.txt
Madison_fed_42.txt
Hamilton_fed_76.txt
dispt_fed_55.txt
Hamilton_fed_1.txt
Hamilton_fed_80.txt
Hamilton_fed_9.txt
Hamilton_fed_12.txt
Hamilton_fed_8.txt
Hamilton_fed_19.txt
Hamilton_fed_32.txt
Hamilton_fed_75.txt
Hamilton_fed_77.txt
Hamilton_fed_21.txt
Hamilton_fed_69.txt
Hamilton_fed_65.txt
Hamilton_fed_66.txt
Hamilton_fed_59.txt
Hamilton_fed_60.txt
Hamilton_fed_30.txt
Hamilton_fed_34.txt
Hamilton_fed_33.txt
Hamilton_fed_83.txt
Hamilton_fed_84.txt
Hamilton_fed_78.txt
Hamilton_fed_81.txt
Hamilton_fed_85.txt
Hamilton_fed_15.txt
Hamilton_fed_68.txt
Hamilton_fed_29.txt
Hamilton_fed_26.txt
Hamilton_fed_24.txt
Hamilton_fed_73.txt
Hamilton_fed_36.txt
Hamilton_fed_71.txt
Hamilton_fed_72.txt
Hamilton_fed_27.txt
Hamilton_fed_82.txt
Hamilton_fed_23.txt
Hamilton_fed_67.txt
dispt_fed_51.txt
dispt_fed_54.txt
Madison_fed_37.txt
dispt_fed_57.txt
Madison_fed_14.txt
dispt_fed_49.txt
dispt_fed_52.txt
Madison_fed_58.txt
Madison_fed_43.txt
dispt_fed_63.txt
Madison_fed_41.txt
dispt_fed_56.txt
Madison_fed_10.txt
dispt_fed_50.txt
dispt_fed_62.txt
Madison_fed_45.txt
Madison_fed_46.txt
Madison_fed_44.txt
Madison_fed_47.txt
dispt_fed_20.txt
HM_fed_18.txt
HM_fed_19.txt
Jay_fed_2.txt
Jay_fed_64.txt
Jay_fed_5.txt
Jay_fed_3.txt
Jay_fed_4.txt
distance3
hclust (*, "complete")