ISSR Short Course

Gregory J. Matthews [1]

[1]Department of Mathematics and Statistics
Loyola University Chicago

June 2015

Outline

Jaccard Index

Similarity

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Dissimilarity

$$J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

```r
library(tm)

## Loading required package:  NLP

#Example of jaccard index
example<-list()
example[[1]]<-PlainTextDocument("The cat dog dog")
example[[2]]<-PlainTextDocument("The dog")
exampleCorp<-Corpus(VectorSource(example))
exampleTDM<-TermDocumentMatrix(exampleCorp)
proxy::dist(as.matrix(t(exampleTDM)), method = "Jaccard")

##           1
## 2 0.3333333
```

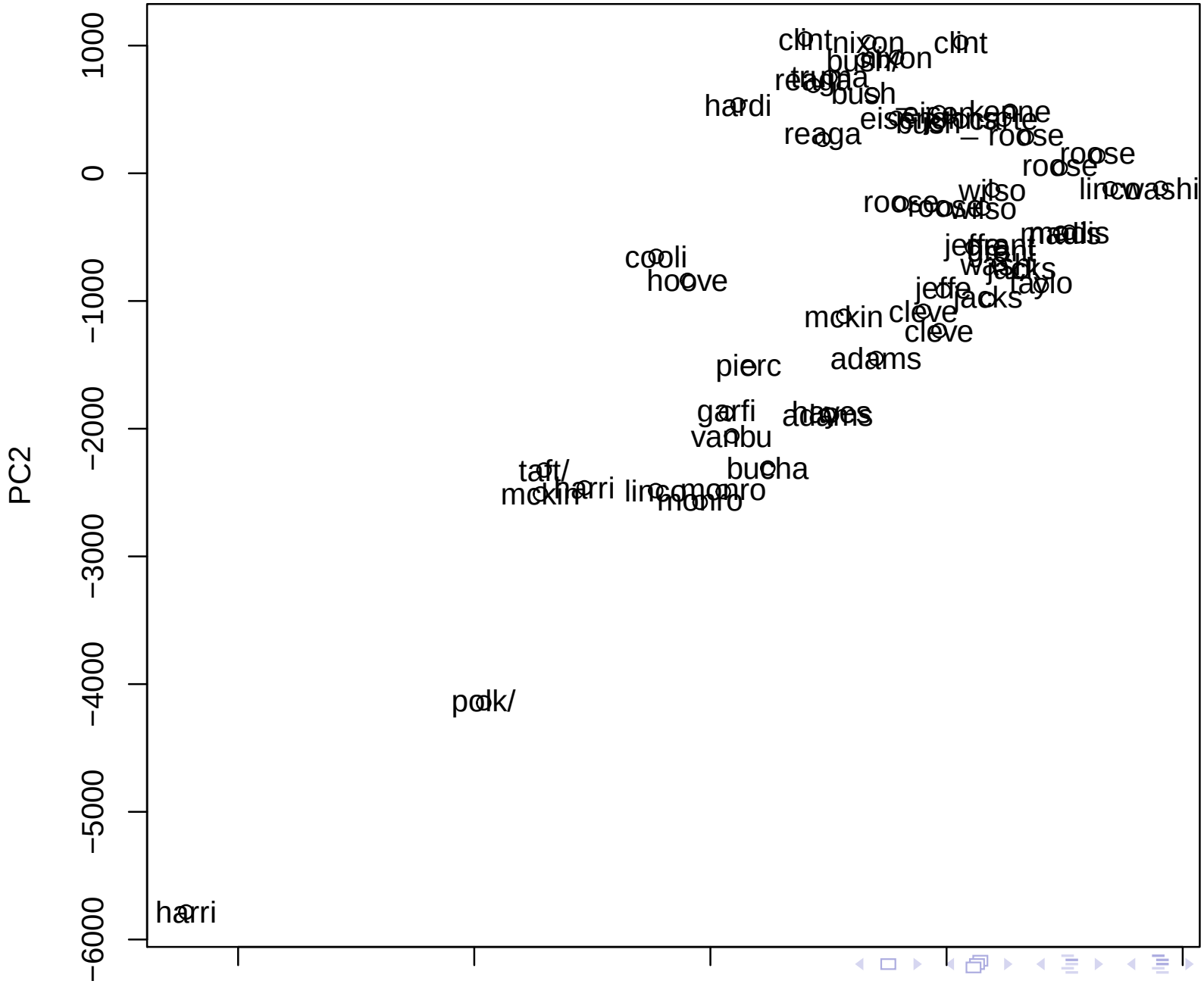
```r
x<-c("The","cat","dog","dog")
y<-c("The","dog")
#Jaccard Index
1-length(intersect(x,y))/length(union(x,y))

## [1] 0.3333333
```

```r
library(tm)
#Example of jaccard index
example<-list()
example[[1]]<-PlainTextDocument("The cat dog dog")
example[[2]]<-PlainTextDocument("The dog")
exampleCorp<-Corpus(VectorSource(example))
exampleTDM<-TermDocumentMatrix(exampleCorp)
proxy::dist(as.matrix(t(exampleTDM)), method = "eJaccard")

##     1
## 2 0.4
```

```r
x<-c(1,2,1)
y<-c(0,1,1)
#eJaccard Index
1-x%*%y/(x%*%x+y%*%y-x%*%y)

##      [,1]
## [1,]  0.4
```

- Principal Component Analysis (PCA) is a dimensionality reduction technique.

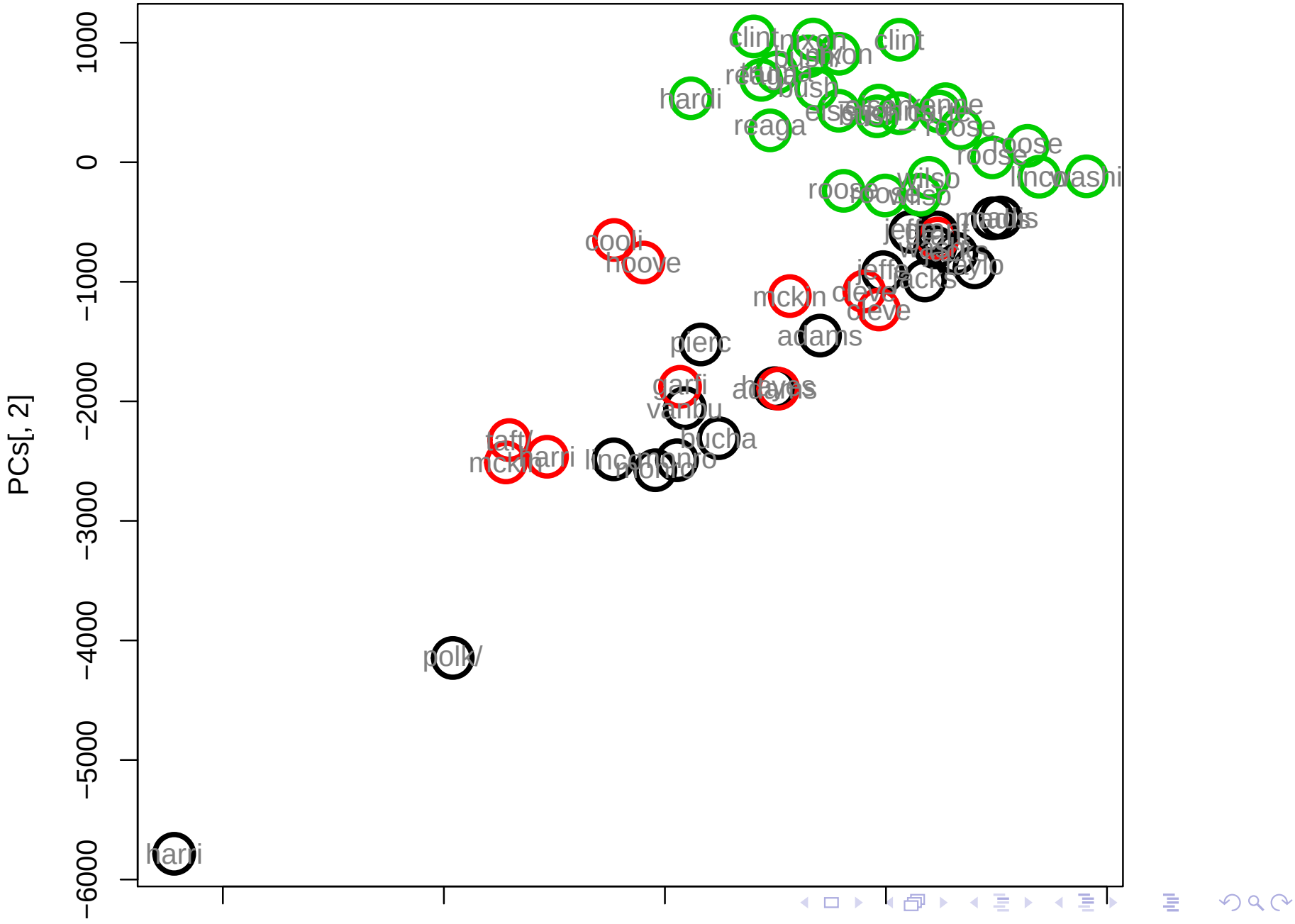
```r
#Principal Component Analysis
corr<-cor(as.matrix(presTDM))
evv<-eigen(corr)
pcs <- evv$vectors[,1:2]
##values:a vector containing the p eigenvalues of x, sorted in decreasing order
evals <- evv$values[1:2]
temp<-diag(1/sqrt(evals)) %*% t(pcs)%*%t(as.matrix(presTDM))
PCs<-t(as.matrix(presTDM))%*%t(temp)
```


PC1 vs PC2: Pres Inaugural Speeches
PC2
1000
0
−1000
−2000
−3000
−4000
−5000
−6000

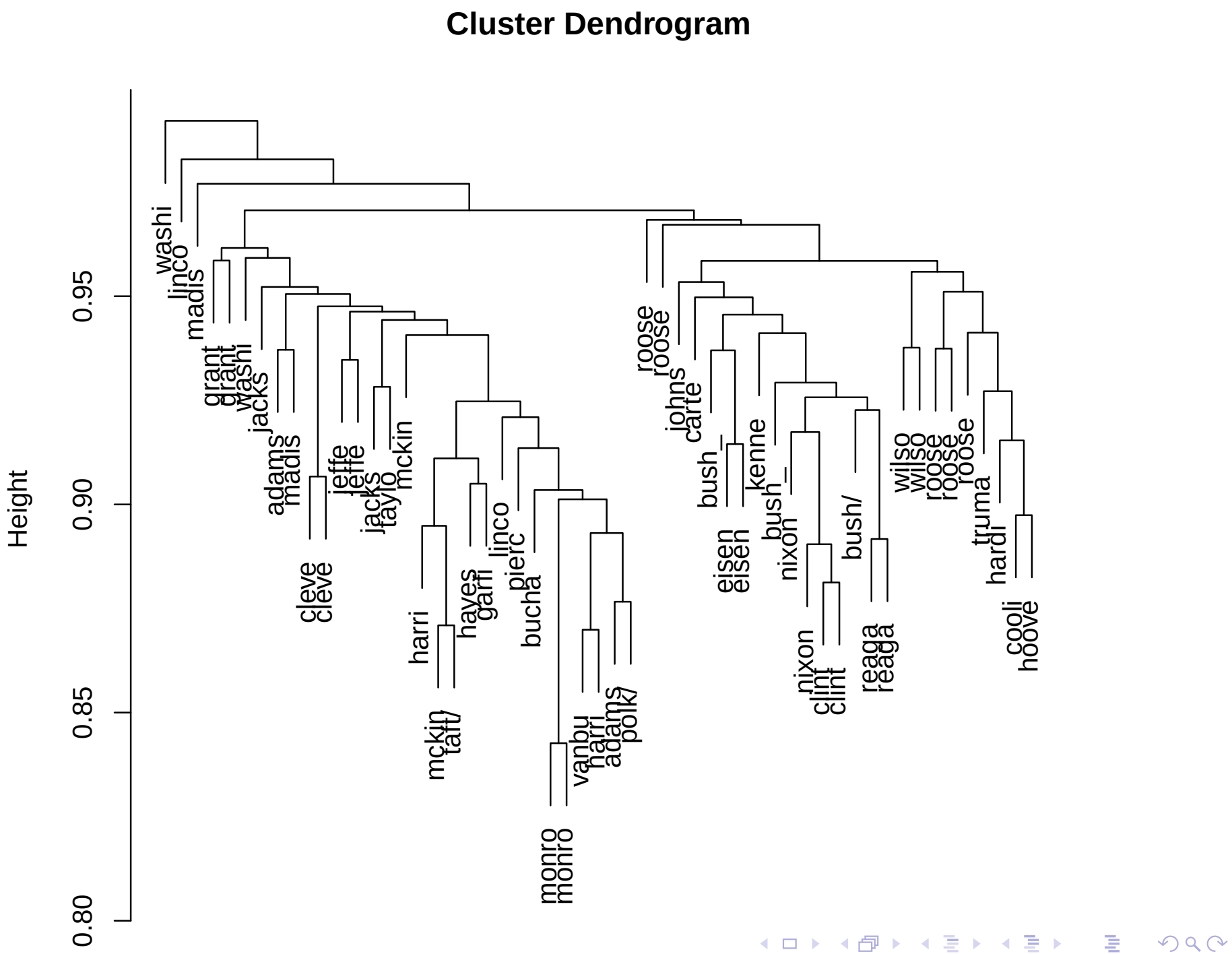
- Cluster analysis allows us to cluster objects into one of K clusters where K can be either known or unknown.
- We'll look at a simple clustering method called k-means.
- Essentially, clusters are created and the center of the cluster is calculated.
- Objects end up in the cluster so that the distances between the object and the cluster centroid are minimized.

- Let's go back to the presidential inaugurual speeches and cluster those.
- We will start with 2 clusters and then explore the possiblity of more clusters.

```r
#Term frequency - Inverse document frequency
m<-weightTfIdf(presTDM)
d<-proxy::dist(as.matrix(t(m)), method = "eJaccard")
cl <- kmeans(d, 3)
table(cl$cluster)

##
##  1  2  3
## 19 11 25
```


PC1 vs PC2: k−means clustering
PCs[, 2]
1000
0
−1000
−2000
−3000
−4000
−5000
−6000
clint
reaga
hardi
roose
lincoln
cooli
hoove
mckin
adams
pierc
polk
harri

```
m<-weightTfIdf(presTDM)
d<-proxy::dist(as.matrix(t(m)), method = "eJaccard")
hc <- hclust(d, method="average")
```


Cluster Dendrogram
Height
0.80
0.85
0.90
0.95
washi
linco
madis
grant
grant
washi
jacks
adams
madis
cleve
cleve
jeffe
jeffe
jacks
taylo
mckin
harri
mckin
taft
haves
garfi
linco
pierc
bucha
monro
monro
vanbu
harri
adams
polk
roose
roose
johns
carte
bush_
eisen
eisen
kenne
bush_
nixon
nixon
clint
clint
bush/
reaga
reaga
wilso
wilso
roose
roose
roose
truma
hardi
cooli
hoove