EC994: Hierarchical Clustering

Thiemo Fetzer

University of Warwick

February 21, 2017

Hierarchical Clustering

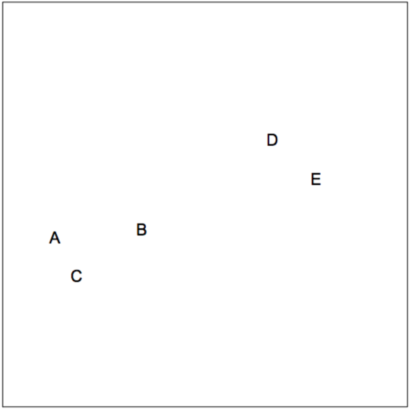
- k-Means clustering is a very powerful tool to detect patterns some data matrix X. However, it has some drawbacks.
- **Numeric features**: In the construction, we compute Euclidian distances between observations and their distinct means - we implicitly assume that the data are *numeric* → k-medoids can fix this!
- **Choice of K**: We are required to make a choice over the number of clusters in the data.
- Hierarchical clustering is a powerful approach that can build clusters based on data dissimilarity matrices and does not require a choice of clusters K.

Hierarchical Clustering Idea

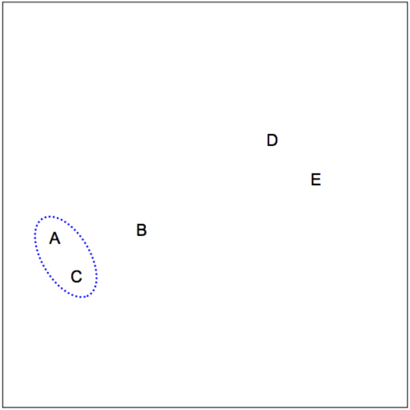
We will discuss one form of hierarchical clustering in this course: Agglomerative hierarchical clustering.

- Start with each point in its own cluster.
- Identify the *closest* two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

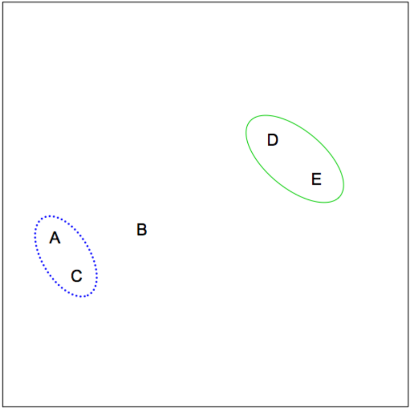
Hierarchical Clustering Illustration



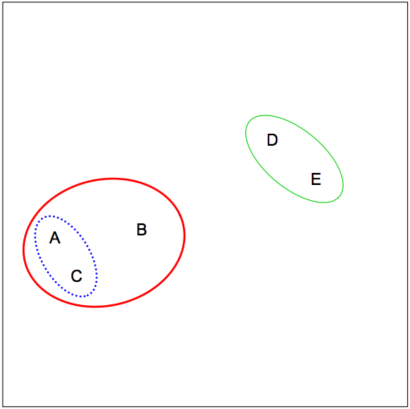
Hierarchical Clustering Illustration



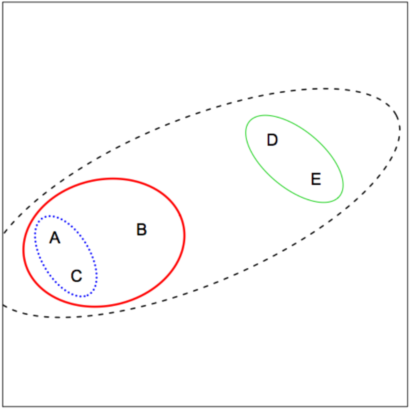
Hierarchical Clustering Illustration



Hierarchical Clustering Illustration

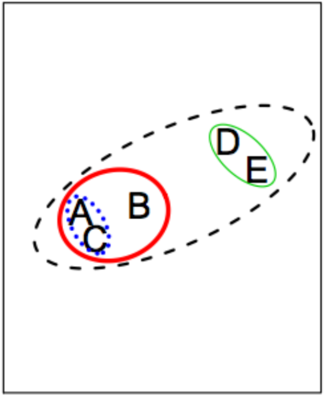


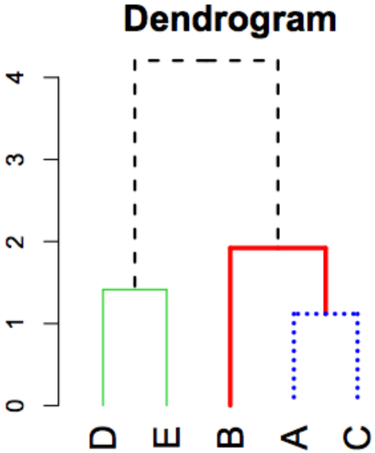
Hierarchical Clustering Illustration



Representing Hierarchical Clustering as Tree

We can not plot out hierarchical clustering results in two dimensions only if there are two dimensional features. With more dimensions, it easiest to display the clustering results in form of a dendogram.



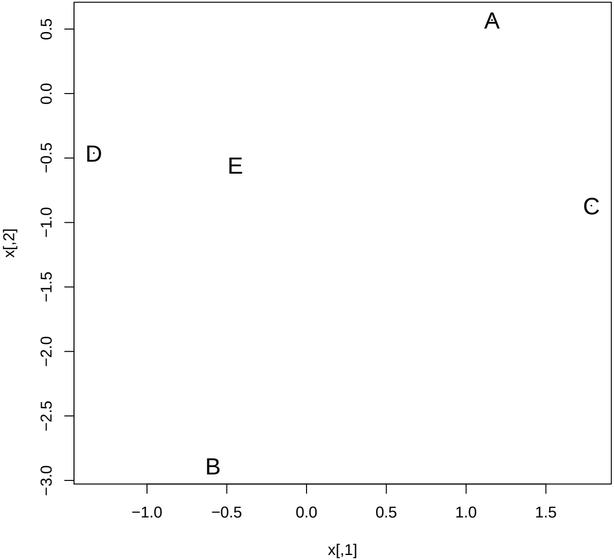


Hierarchical Clustering Algorithm

Algorithm (*Hierarchical Clustering*)

1. *Start by considering each item as its own cluster, for n clusters and calculate the $n(n1)/2$ pairwise distances between each of the n clusters.*
   For $i = n, n-1, ..., 2$
2. *Examine all pair-wise inter-cluster dissimilarities among the i clusters, identify the pair of clusters k, k' that are least dissimilar and combine them. Store the distance betweent k and k'.*
3. *Recalculate distance matrix with the remaining $i-1$ cluster.*
4. *Stop when only two clusters remain.*

Another example performing hierarchical clustering



Another example performing hierarchical clustering

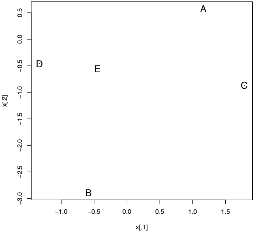
```
##        A    B    C     D     E
## A 0.00 3.88 1.57 2.700 1.964
## B 3.88 0.00 3.12 2.540 2.338
## C 1.57 3.12 0.00 3.145 2.254
## D 2.70 2.54 3.14 0.000 0.891
## E 1.96 2.34 2.25 0.891 0.000
```

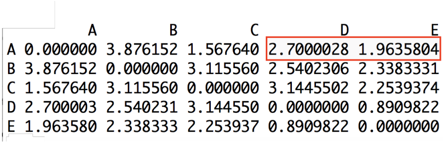
Sequence of Steps

- We start with every observation being its own cluster, i.e. we are at the bottom of the tree and there are five clusters.
- In the first iteration, we combine combine D and E. Their distance is 0.891, so we record this distance. Now we are left with four clusters DE, A, B, C.
- The second step in the algorithm asks us to recompute the distance matrix, but how should we do this?
- There are multiple choices.

Different Linkage choices: Complete Linkage

Complete Linkage: This maximizes intercluster dissimilarity, as we compute all pairwise distances between the observations in cluster A and the observations in cluster B and record the *largest* in our reduced distance matrix.

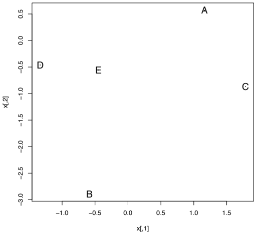


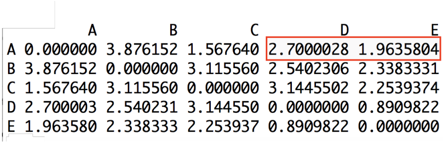


```
         A        B        C         D         E
A 0.000000 3.876152 1.567640 2.7000028 1.9635804
B 3.876152 0.000000 3.115560 2.5402306 2.3383331
C 1.567640 3.115560 0.000000 3.1445502 2.2539374
D 2.700003 2.540231 3.144550 0.0000000 0.8909822
E 1.963580 2.338333 2.253937 0.8909822 0.0000000
```

Different Linkage choices: Single Linkage

Single Linkage: This leads to minimal intercluster dissimilarity, as we compute all pairwise distances between the observations in cluster A and the observations in cluster B and record the *minimal* in our reduced distance matrix.

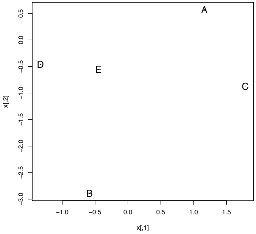


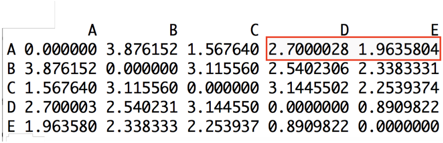


```
          A        B        C         D         E
A 0.000000 3.876152 1.567640 2.7000028 1.9635804
B 3.876152 0.000000 3.115560 2.5402306 2.3383331
C 1.567640 3.115560 0.000000 3.1445502 2.2539374
D 2.700003 2.540231 3.144550 0.0000000 0.8909822
E 1.963580 2.338333 2.253937 0.8909822 0.0000000
```

Different Linkage choices: Average Linkage

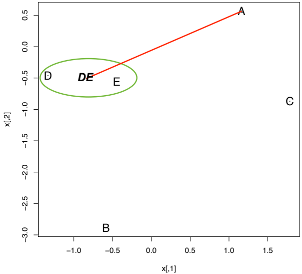
Average Linkage: Compute all pairwise dissimilarities between observations in cluster A and observations in cluster B and record the average of these dissimilarities.

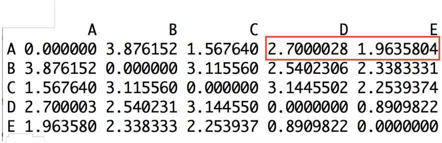



```
         A        B        C         D         E
A 0.000000 3.876152 1.567640 2.7000028 1.9635804
B 3.876152 0.000000 3.115560 2.5402306 2.3383331
C 1.567640 3.115560 0.000000 3.1445502 2.2539374
D 2.700003 2.540231 3.144550 0.0000000 0.8909822
E 1.963580 2.338333 2.253937 0.8909822 0.0000000
```

Different Linkage choices: Centroid Linkage

Centroid Linkage: Compute the centroid within each cluster.

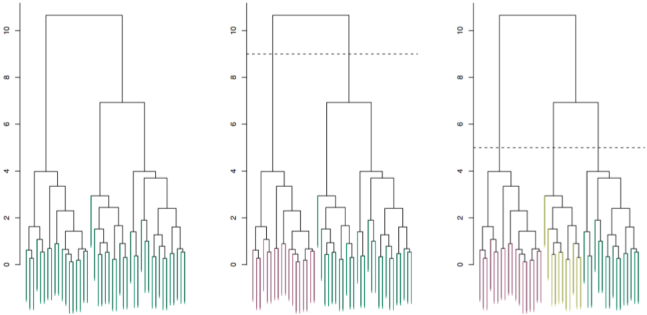




```
         A        B        C         D         E
A 0.000000 3.876152 1.567640 2.7000028 1.9635804
B 3.876152 0.000000 3.115560 2.5402306 2.3383331
C 1.567640 3.115560 0.000000 3.1445502 2.2539374
D 2.700003 2.540231 3.144550 0.0000000 0.8909822
E 1.963580 2.338333 2.253937 0.8909822 0.0000000
```

Number of Clusters

By cutting at different heights of the tree, you can control the number of clusters to consider.



Correlation based distance metrics

- Rather than caring about levels of distance, we may be interested in correlation based distances; for example, a point described as $(1, 1)$ may be quite far away from a point $(10, 10)$, but they are actually perfectly correlated if we think of them as vectors.
- A correlation measure would take this scaling factor out and indicate that these two vectors are identical!
- In many instances, we care about pattern correlatedness and not necessarily pure distance.
- Last week, we have already seen one example

Cosine Similarity: Measuring Angle between two unit length vectors

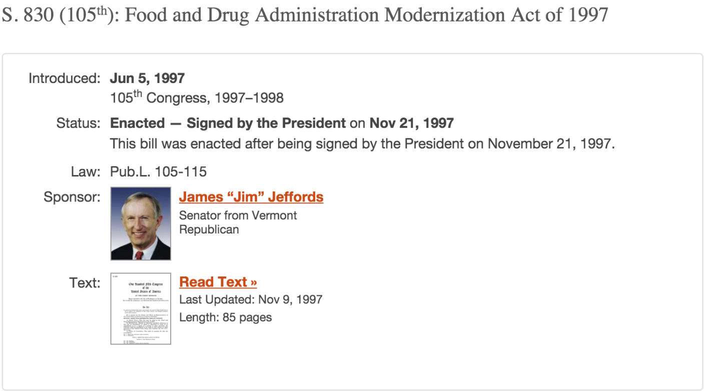
- What is the length of a vector A and B? its simply the Euclidian distance from origin, i.e. $\|\mathbf{y_A}\|$,$\|\mathbf{y_B}\|$
- So the vectors $\mathbf{y'_A} = \frac{\mathbf{y_A}}{\|\mathbf{y_A}\|}$ and $\mathbf{y'_B} = \frac{\mathbf{y_B}}{\|\mathbf{y_B}\|}$ both have length 1.
- What is the angle between the vectors $\frac{\mathbf{y_A}}{\|\mathbf{y_A}\|}$ and $\frac{\mathbf{y_B}}{\|\mathbf{y_B}\|}$?

$$\cos\left(\mathbf{y_A}, \mathbf{y_B}\right) = \frac{\mathbf{y_A} \cdot \mathbf{y_B}}{\|\mathbf{y_A}\|\|\mathbf{y_B}\|} = \frac{\sum_{i=1}^{n} y_{iA} y_{iB}}{\sqrt{\sum_{i=1}^{n} y_{iA}^2}\sqrt{\sum_{i=1}^{n} y_{iB}^2}}$$

Mapping Legislative Influence

- Most proposed bills do never make it into actual law.
- Consider the following examples for the US:
- H.R. 1060 (105th): Pharmacy Compounding Act
- S. 830 (105th): Food and Drug Administration Modernization Act of 1997

Mapping Legislative Influence



S. 830 (105th): Food and Drug Administration Modernization Act of 1997

Introduced: **Jun 5, 1997**
105th Congress, 1997–1998

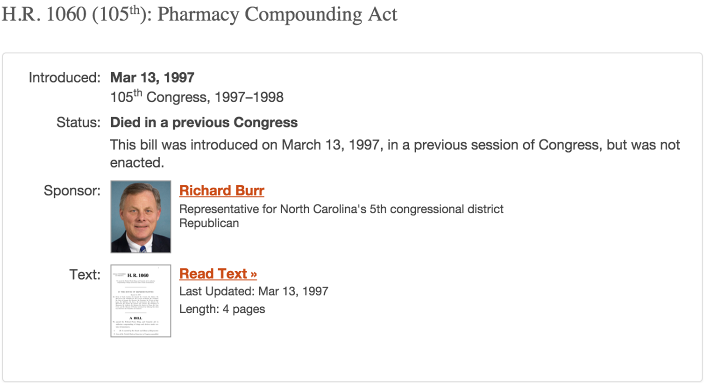
Status: **Enacted — Signed by the President** on **Nov 21, 1997**
This bill was enacted after being signed by the President on November 21, 1997.

Law: Pub.L. 105-115

Sponsor: **James “Jim” Jeffords**
Senator from Vermont
Republican

Text: **Read Text »**
Last Updated: Nov 9, 1997
Length: 85 pages

Mapping Legislative Influence



H.R. 1060 (105th): Pharmacy Compounding Act

Introduced: **Mar 13, 1997**
105th Congress, 1997–1998

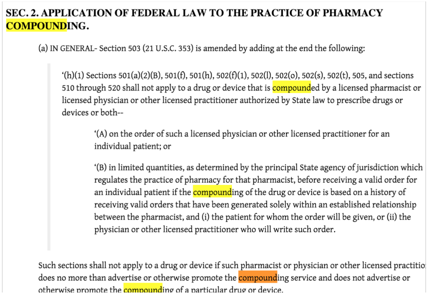
Status: **Died in a previous Congress**

This bill was introduced on March 13, 1997, in a previous session of Congress, but was not enacted.

Sponsor: **Richard Burr**
Representative for North Carolina's 5th congressional district
Republican

Text: **Read Text »**
Last Updated: Mar 13, 1997
Length: 4 pages

Mapping Legislative Influence

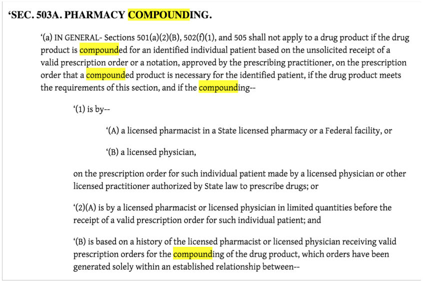


SEC. 2. APPLICATION OF FEDERAL LAW TO THE PRACTICE OF PHARMACY COMPOUNDING.

(a) IN GENERAL- Section 503 (21 U.S.C. 353) is amended by adding at the end the following:

> '(h)(1) Sections 501(a)(2)(B), 501(f), 501(h), 502(f)(1), 502(l), 502(o), 502(s), 502(t), 505, and sections 510 through 520 shall not apply to a drug or device that is compounded by a licensed pharmacist or licensed physician or other licensed practitioner authorized by State law to prescribe drugs or devices or both--
>
> '(A) on the order of such a licensed physician or other licensed practitioner for an individual patient; or
>
> '(B) in limited quantities, as determined by the principal State agency of jurisdiction which regulates the practice of pharmacy for that pharmacist, before receiving a valid order for an individual patient if the compounding of the drug or device is based on a history of receiving valid orders that have been generated solely within an established relationship between the pharmacist, and (i) the patient for whom the order will be given, or (ii) the physician or other licensed practitioner who will write such order.

Such sections shall not apply to a drug or device if such pharmacist or physician or other licensed practitio does no more than advertise or otherwise promote the compounding service and does not advertise or otherwise promote the compounding of a particular drug or device.



'SEC. 503A. PHARMACY COMPOUNDING.

'(a) IN GENERAL- Sections 501(a)(2)(B), 502(f)(1), and 505 shall not apply to a drug product if the drug product is compounded for an identified individual patient based on the unsolicited receipt of a valid prescription order or a notation, approved by the prescribing practitioner, on the prescription order that a compounded product is necessary for the identified patient, if the drug product meets the requirements of this section, and if the compounding--

'(1) is by--

'(A) a licensed pharmacist in a State licensed pharmacy or a Federal facility, or

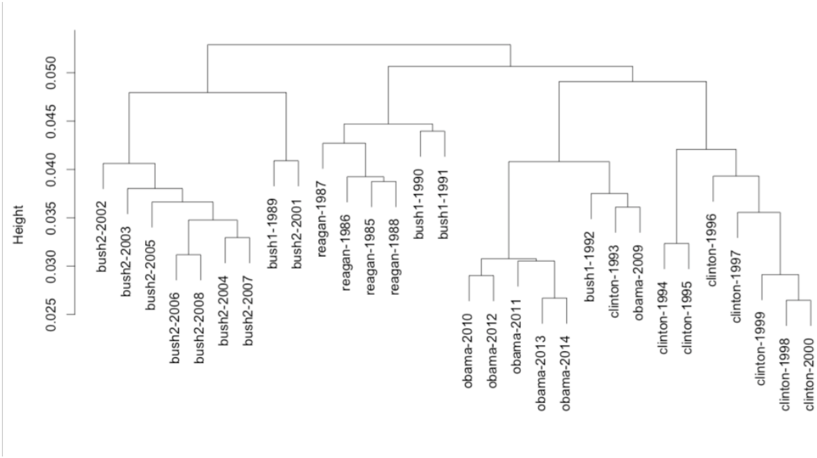
'(B) a licensed physician,

on the prescription order for such individual patient made by a licensed physician or other licensed practitioner authorized by State law to prescribe drugs; or

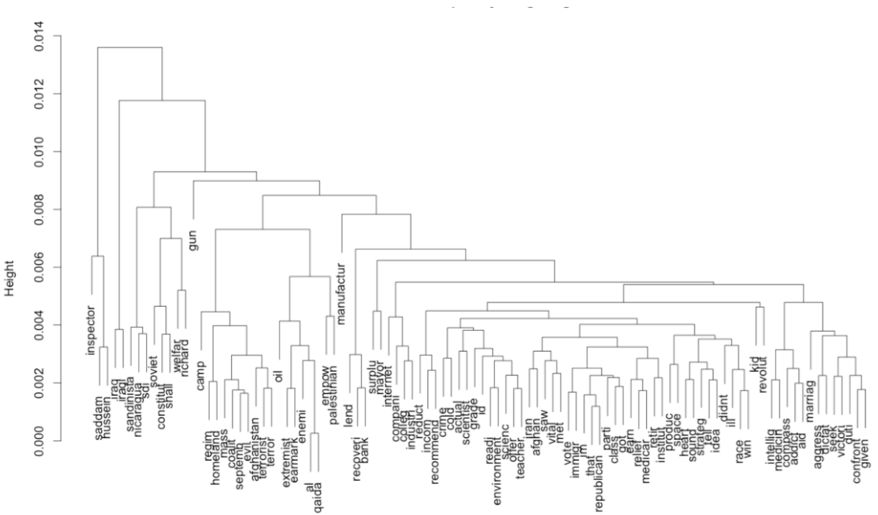
'(2)(A) is by a licensed pharmacist or licensed physician in limited quantities before the receipt of a valid prescription order for such individual patient; and

'(B) is based on a history of the licensed pharmacist or licensed physician receiving valid prescription orders for the compounding of the drug product, which orders have been generated solely within an established relationship between--

State of the Union Clustering



Word Feature Clustering to identify collocated words



Some hands-on R code on clustering

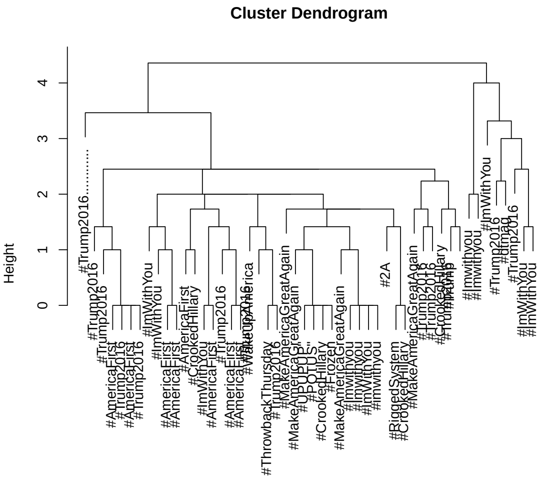
```r
library(cluster)
library(quanteda)


## quanteda version 0.9.9.3
##
## Attaching package:  ’quanteda’
## The following object is masked from ’package:utils’:
##
##     View
## The following object is masked from ’package:base’:
##
##     sample

load(file = "../../Data/trumpstweets.rdata")
# remove retweet entities remove html links
tw.user.df$text <- gsub("http([^ ]*)", "", tw.user.df$text)
tw.user.df[, `:=`(firsthash, str_extract(text, "#([^ ]*)"))]
tw.user.df <- tw.user.df[!is.na(firsthash)]

tw.user.df$text <- gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tw.user.df$text)
# remove at people
tw.user.df$text <- gsub("@\\w+", "", tw.user.df$text)
# remove punctuation
tw.user.df$text <- gsub("[[:punct:]]", "", tw.user.df$text)
# remove numbers
tw.user.df$text <- gsub("[[:digit:]]", "", tw.user.df$text)
# remove unnecessary spaces
tw.user.df$text <- gsub("[ \t]{2,}", "", tw.user.df$text)
tw.user.df$text <- gsub("^\\s+|\\s+$", "", tw.user.df$text)
tw.user.df[, `:=`(docname, paste("text", 1:nrow(tw.user.df), sep = ""))]
trump.dfm <- dfm(tw.user.df$text, remove = stopwords(), stem = TRUE)
```

Hierarchical Clustering on Trump tweets

```r
## k-means clustering

set.seed(18022017)
trump.dfm.trim <- dfm_trim(trump.dfm, min_count = 25, min_docfreq = 10)


## Removing features occurring:
##   - fewer than 25 times:  1,376
##   - in fewer than 10 documents:  1,328
##   Total features removed:  1,376 (98.1%).

trump.dfm.trim <- trump.dfm.trim[1:50, ]
# tf function converts word count dfm to share
trump.dfm.dist <- dist(as.matrix(dfm_weight(trump.dfm.trim, "frequency")))

trump.dfm.clust <- hclust(trump.dfm.dist)
# label with document names
trump.dfm.clust$labels <- tw.user.df[docname %in% docnames(trump.dfm.trim)]$firsthash
```

Hierarchical Clustering on Trump tweets

```r
## k-means clustering plot as a dendrogram
plot(trump.dfm.clust, xlab = "", sub = "")
```



Hierarchical Clusterong on SOTUs

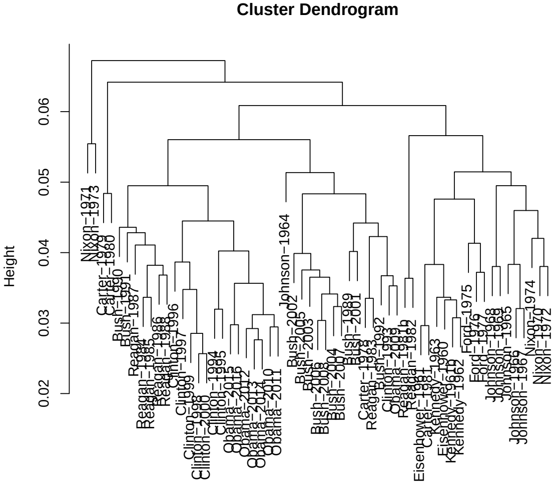
```
## k-means clustering
data(SOTUCorpus, package = "quantedaData")
presDfm <- dfm(corpus_subset(SOTUCorpus, Date > as.Date("1960-01-01")), verbose = FALSE, stem = TRUE,
    remove = stopwords("english"), removePunct = TRUE)
presDfm <- dfm_trim(presDfm, min_count = 5, min_docfreq = 3)
```

```
## Removing features occurring:
##   - fewer than 5 times:  5,857
##   - in fewer than 3 documents:  5,115
##   Total features removed:  5,908 (64.7%).
```

```
# hierarchical clustering - get distances on normalized dfm
presDistMat <- dist(as.matrix(dfm_weight(presDfm, "relFreq")))
# hiarchical clustering the distance object
presCluster <- hclust(presDistMat)
# label with document names
presCluster$labels <- docnames(presDfm)
# plot as a dendrogram
```

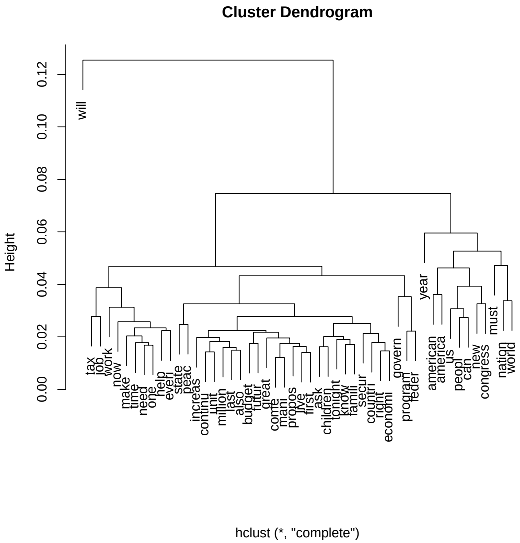
Hierarchical Clusterong on SOTUs

```
plot(presCluster, xlab = "", sub = "")
```



Identify Bigrams

```r
wordDfm <- dfm_sort(dfm_weight(presDfm, "relFreq"))  # sort in decreasing order of total word freq
wordDfm <- t(wordDfm)[1:50, ]  # because transposed
wordDistMat <- dist(wordDfm)
wordCluster <- hclust(wordDistMat)
plot(wordCluster, xlab = "")
```


Cluster Dendrogram
Height
will
tax
job
work
now
make
time
need
one
help
everi
state
peac
increas
continu
unit
million
last
also
budget
futur
great
come
mani
propos
live
first
ask
children
tonight
know
famili
secur
countri
right
economi
govern
program
feder
year
american
america
us
peopl
can
new
congress
must
nation
world
hclust (*, "complete")