2022 SISBID Clustering Lab

Genevera I. Allen & Yufeng Liu

Data set - Author Data.

This data set consists of word counts from chapters written by four authors.

This lab will put together concepts from both dimension reduction and clustering.

There are ultimately 3 goals to this lab:
* Correctly cluster author texts in an unsupervised manner.
* Determine which words are responsible for correctly separating the author texts.
* Visualize the author texts, words and the results of your analysis.

1. Problem 1 - Visualization

- Problem 1a - We wish to plot the author texts as well as the words via a 2D scatterplot. Which method would be best to use? Why?
- Problem 1b - Apply PCA to visualize the author texts. Explain the results.
- Problem 1c - Apply MDS to visualize the author texts. Interpret the results.
- Problem 1d - Can you use MDS to help determine which distance is appropriate for this data? Which one is best and why?
- Problem 1e - Apply MDS with your chosen distance to visualize the words. Interpret the results.

2. Problem 2 - K-means

- Problem 2a - Apply K-means with K=4 to this data.
- Problem 2b - How well does K-mean do at separating the authors?
- Problem 2c - Is K-means an appropriate clustering algorithm for this data? Why or Why not?

3. Problem 3 - Hierarchical Clustering

- Problem 3a - Apply hierarchical clustering to this data set.
- Problem 3b - Which distance is best to use? Why?
- Problem 3c - Which linkage is best to use? Why?
- Problem 3d - Do any linkages perform particularly poorly? Explain this result.
- Problem 3e - Visualize your hierarchical clustering results.

4. Problem 4 - Biclustering

- Problem 4a - Apply the cluster heatmap method to visualize this data. Which distance and linkage functions did you use?
- Problem 4b - Interpret the cluster heatmap. Which words are important for distinguishing author texts?

5. Problem 5 - NMF

- Problem 5a - Apply NMF with K = 4 and use W to assign cluster labels to each observation.
- Problem 5b - How well does NMF perform? Interpret and explain this result.

- Problem 5c - Can you use the NMF to determine which words are important for distinguishing author texts? How? What did you find?

6. Problem 6 - Wrap-up

- Problem 6a - Overall, which method is the best at clustering the author texts? Why is this the case?
- Problem 6b - Which words are key for distinguishing the author texts? How did you determine these?
- Problem 6c - Overall, which is the best method for providing a visual summary of the data?

R scripts to help out with the Clustering Lab

Don't peek at this if you want to practice coding on your own!!

Load packages

```
library(NMF)
library(ggplot2)
library(umap)
```

Load dataset: Author data

```
load("UnsupL_SISBID_2022.Rdata")
# understand the data a bit
dim(author)
```

```
## [1] 841  70
```

```
colnames(author)
```

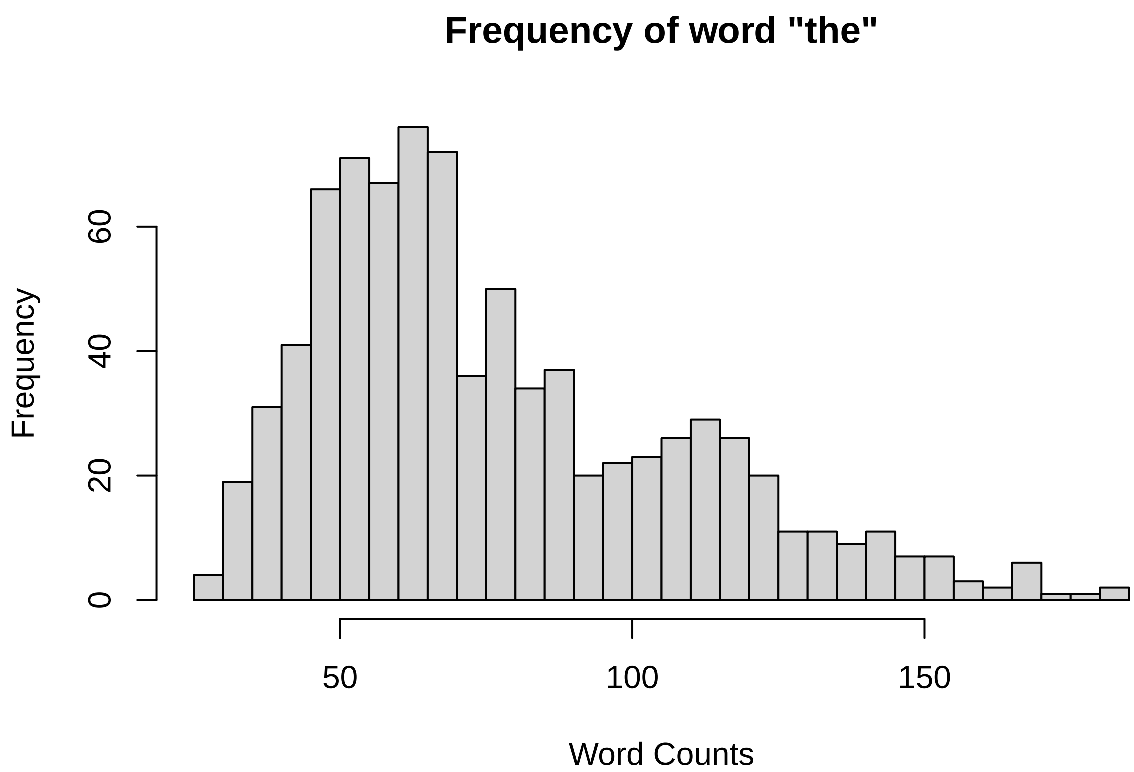
```
##  [1] "a"      "all"    "also"   "an"     "and"    "any"    "are"    "as"
##  [9] "at"     "be"     "been"   "but"    "by"     "can"    "do"     "down"
## [17] "even"   "every"  "for."   "from"   "had"    "has"    "have"   "her"
## [25] "his"    "if."    "in."    "into"   "is"     "it"     "its"    "may"
## [33] "more"   "must"   "my"     "no"     "not"    "now"    "of"     "on"
## [41] "one"    "only"   "or"     "our"    "should" "so"     "some"   "such"
## [49] "than"   "that"   "the"    "their"  "then"   "there"  "things" "this"
## [57] "to"     "up"     "upon"   "was"    "were"   "what"   "when"   "which"
## [65] "who"    "will"   "with"   "would"  "your"   "BookID"
```

```
unique(rownames(author))
```

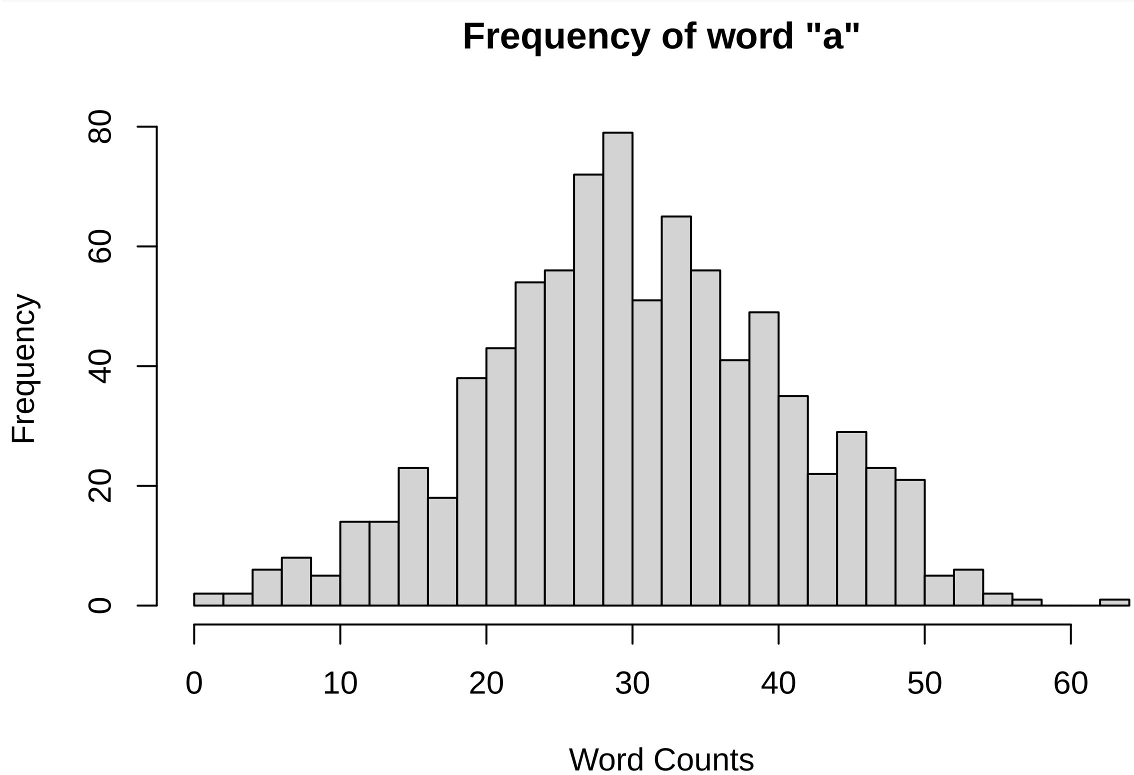
```
## [1] "Austen"      "London"      "Milton"      "Shakespeare"
```

```
TrueAuth = as.factor(rownames(author))
```

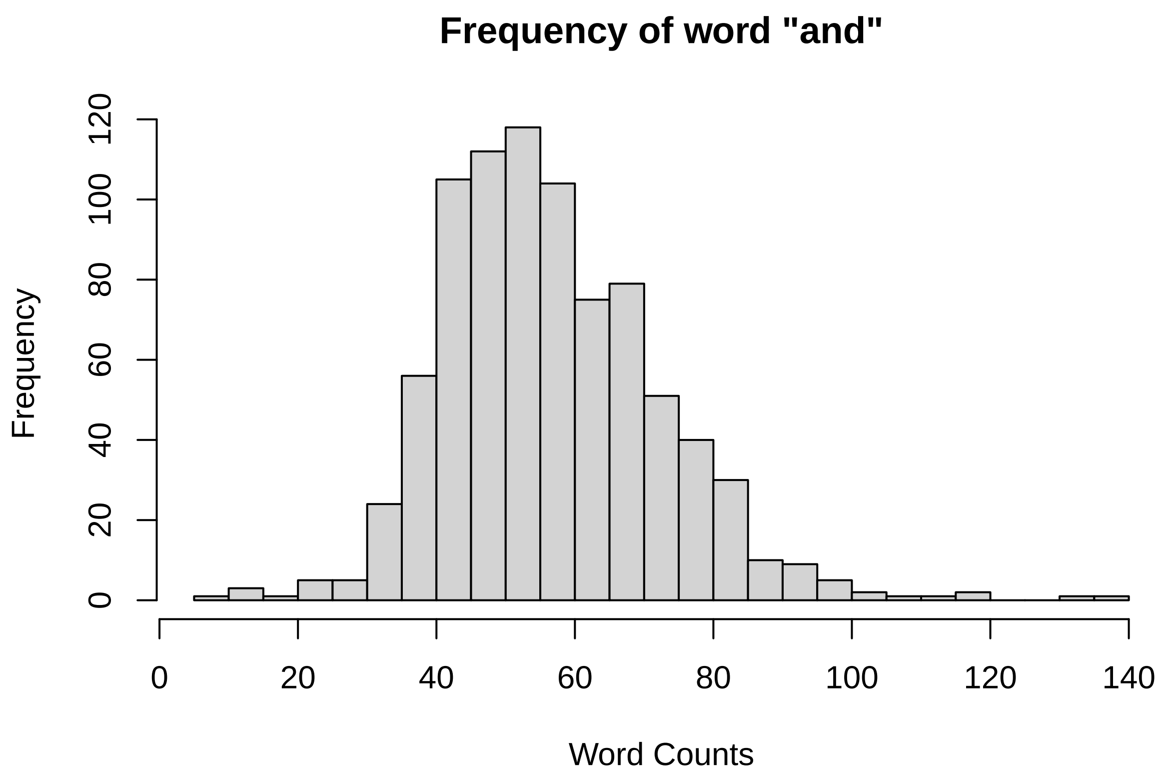
```
hist(author[,colnames(author)=="the"],breaks=25,main="Frequency of word \"the\"",xlab = "Word Counts")
```



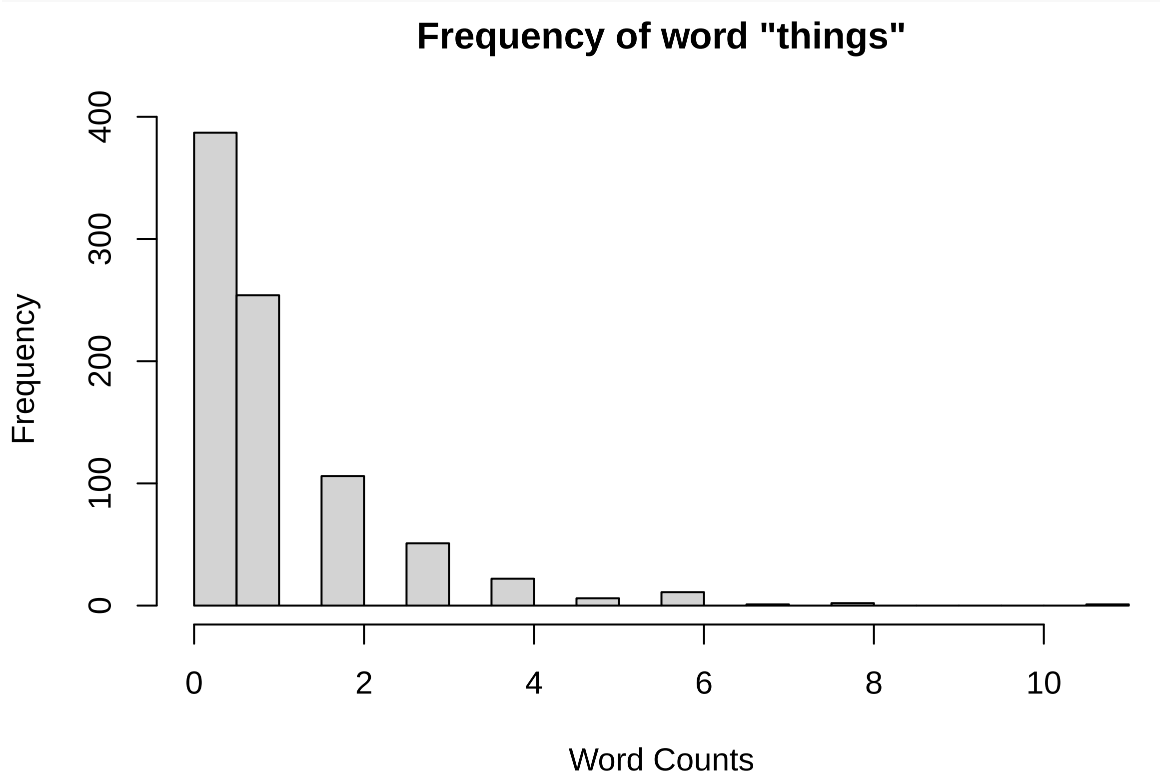

```r
hist(author[,colnames(author)=="a"],breaks=25,main="Frequency of word \"a\"",xlab = "Word Counts")
```




```r
hist(author[,colnames(author)=="and"],breaks=25,main="Frequency of word \"and\"",xlab = "Word Counts")
```




```
hist(author[,colnames(author)=="things"],breaks=25,main="Frequency of word \"things\"",xlab = "Word Cou
```



Take out bookID

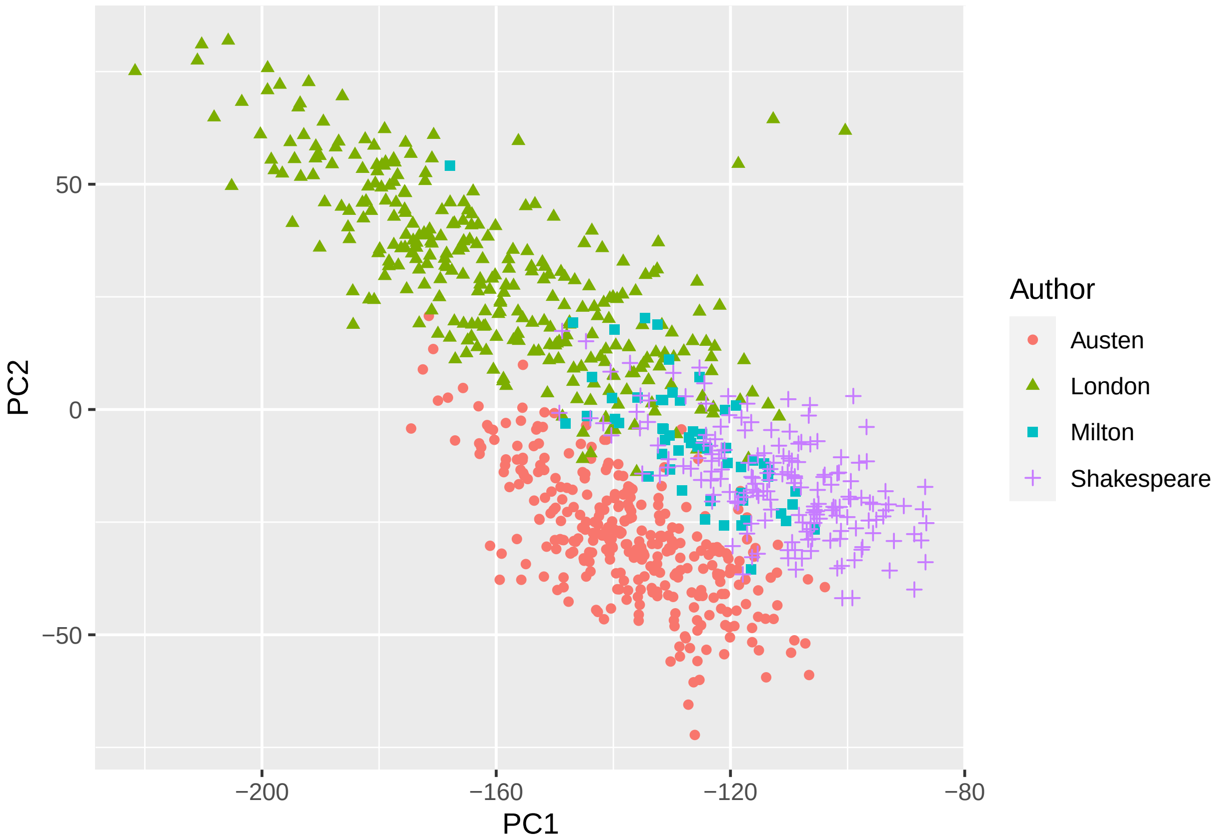
```
AuthorData = author[,1:69]
```

Problem 1 - Visualization

- how to visulaize texts? words? in 2-dimensions

Trying PCA

```
sv = svd(AuthorData)
V = sv$v
Z = AuthorData%*%V
# projected matrix
PCData = data.frame(cbind(Z[,1],Z[,2],rownames(AuthorData)),stringsAsFactors = FALSE)
colnames(PCData) = c("PC1","PC2","Author")
PCData$PC1 = as.numeric(PCData$PC1)
PCData$PC2 = as.numeric(PCData$PC2)
# plot
ggplot(PCData) +
  geom_point(mapping=aes(x = PC1,y= PC2,color = Author,shape= Author))
```



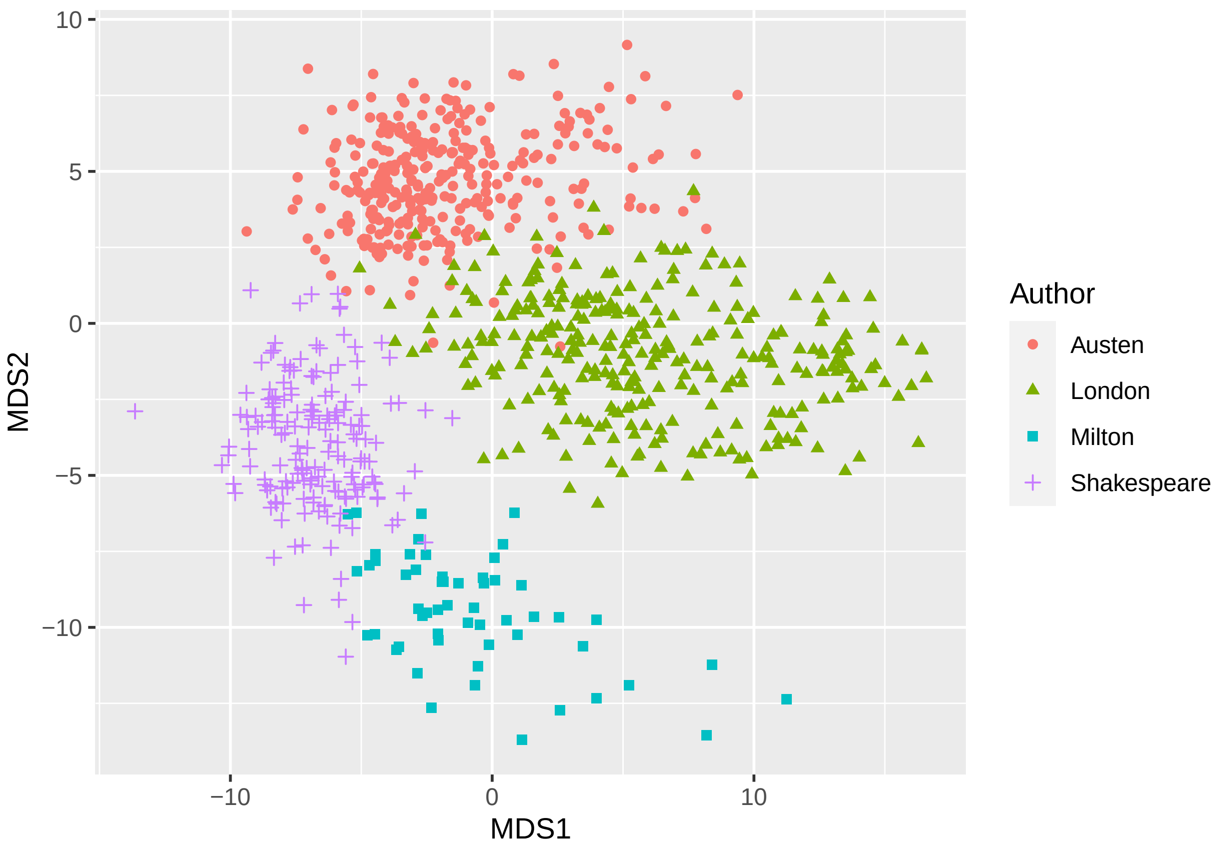
Why doesn't this work well?

Trying MDS (classical)
Can you use MDS to decide which distance is best to understand this data?

Visualizing author texts

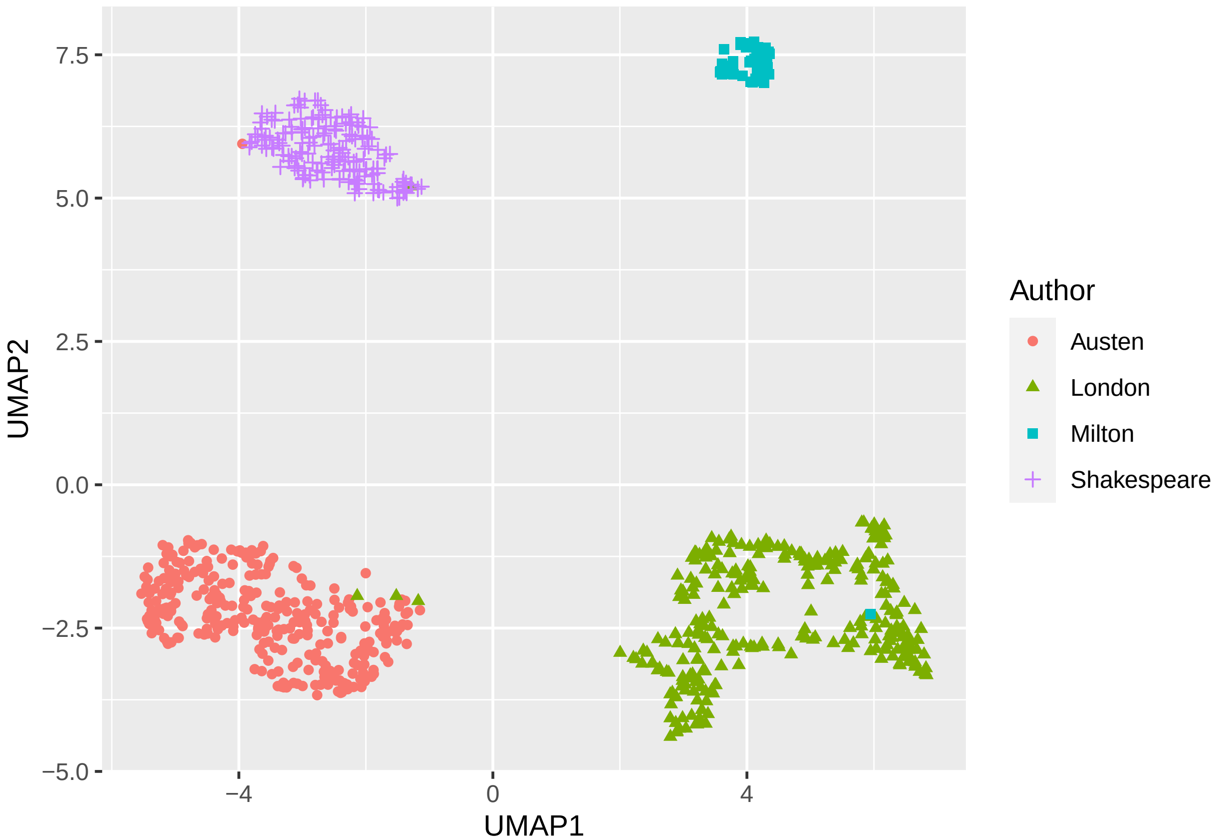
```
Dmat = dist(log(AuthorData+1),method="canberra")
mdsres = cmdscale(Dmat,k=2)
# MDS matrix
MDSData = data.frame(cbind(mdsres[,1],mdsres[,2],rownames(AuthorData)),
                     stringsAsFactors = FALSE)
```

```
colnames(MDSData) = c("MDS1","MDS2","Author")
MDSData$MDS1 = as.numeric(MDSData$MDS1)
MDSData$MDS2 = as.numeric(MDSData$MDS2)
# plot
ggplot(MDSData) +
  geom_point(mapping=aes(x = MDS1,y= MDS2,color = Author,shape= Author))
```



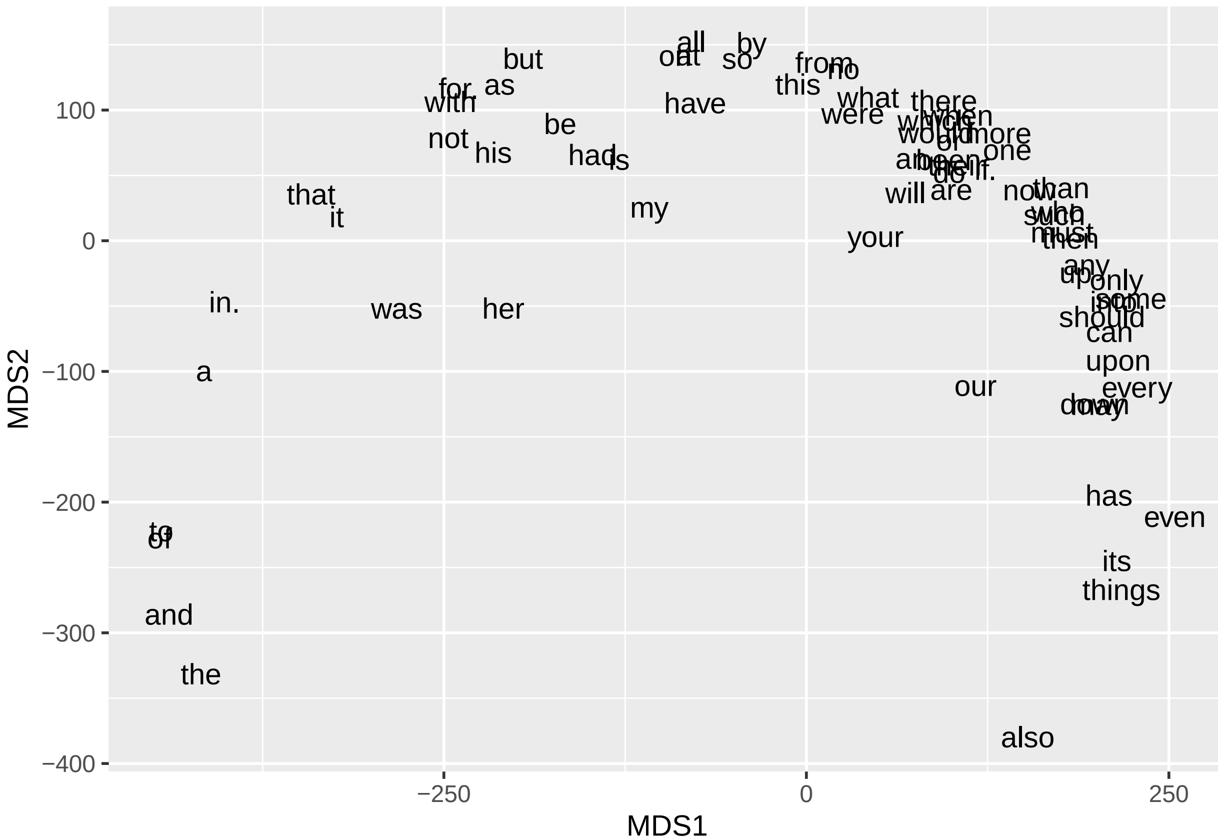
Trying UMAP

```
AuthorUMAP = umap(AuthorData)
# UMAP matrix
UMAPData = data.frame(cbind(AuthorUMAP$layout[,1],AuthorUMAP$layout[,2],
                            rownames(AuthorData)),stringsAsFactors = FALSE)
colnames(UMAPData) = c("UMAP1","UMAP2","Author")
UMAPData$UMAP1 = as.numeric(UMAPData$UMAP1)
UMAPData$UMAP2 = as.numeric(UMAPData$UMAP2)
# plot
ggplot(UMAPData) +
  geom_point(mapping=aes(x = UMAP1,y= UMAP2,color = Author,shape= Author))
```



Visualizing words

```
Dmat = dist(t(AuthorData),method="canberra")
mdsresW = cmdscale(Dmat,k=2)
# MDS matrix for words
MDSDataW = data.frame(cbind(mdsresW[,1],mdsresW[,2],colnames(AuthorData)),
                      stringsAsFactors = FALSE)
colnames(MDSDataW) = c("MDS1","MDS2","Word")
MDSDataW$MDS1 = as.numeric(MDSDataW$MDS1)
MDSDataW$MDS2 = as.numeric(MDSDataW$MDS2)
ggplot(MDSDataW) +
  geom_text(mapping=aes(x = MDS1,y= MDS2,label = Word))
```



Problem 2 - K-means

```
K = 4
km = kmeans(AuthorData,centers=K)
table(km$cluster,TrueAuth)
```

```
##    TrueAuth
##     Austen London Milton Shakespeare
##   1    181     19      0           2
##   2    132     13      0           0
##   3      4    251      2           1
##   4      0     13     53         170
```

Visualization of K-means clustering results via MDS matrix

```
PredData = data.frame(cbind(MDSData[,1:2],km$cluster,rownames(AuthorData)))
colnames(PredData) = c("MDS1","MDS2","PredictedLabel","TrueAuthor")
PredData$PredictedLabel = factor(PredData$PredictedLabel)
ggplot(PredData) +
  geom_point(mapping=aes(x = MDS1,y= MDS2,color = PredictedLabel,shape= TrueAuthor)) +
  ggtitle("Clustering results by K-means with K = 4") +
  theme(plot.title = element_text(hjust = 0.5))
```

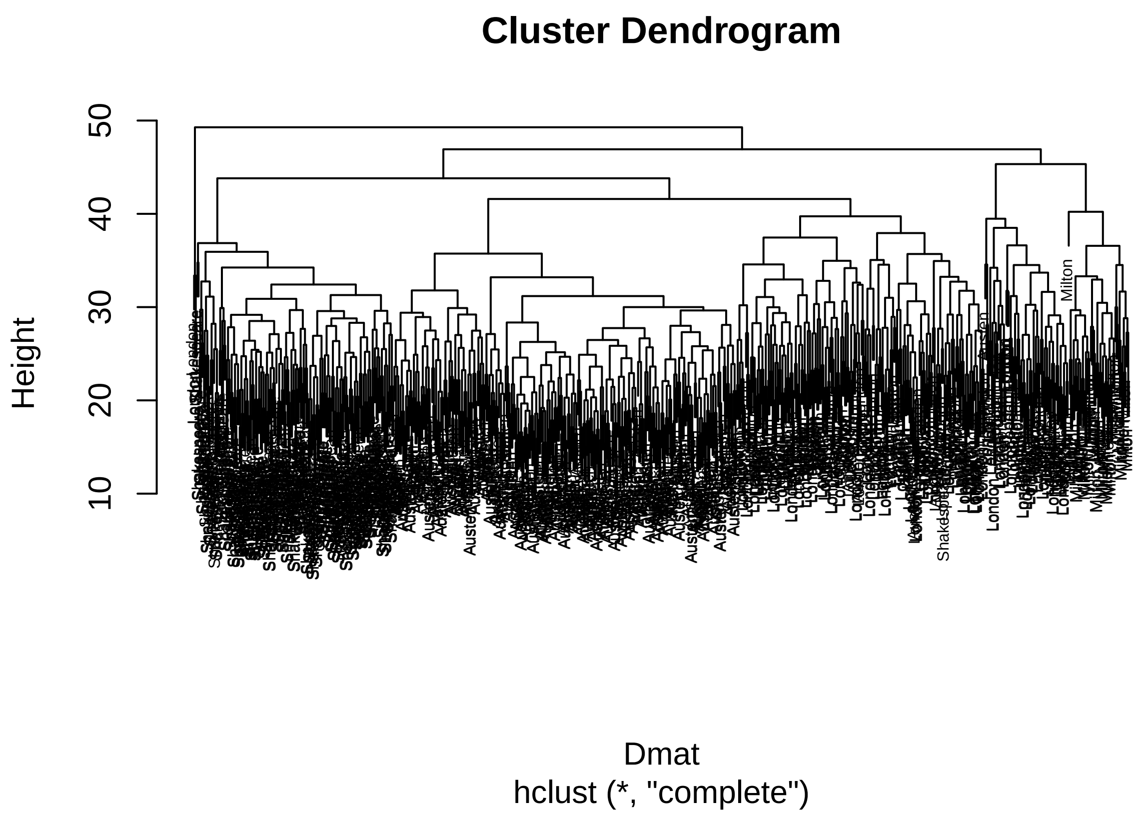
Problem 3 - Hierarchical Clustering

Which distance is appropraite? Why?
canberra distance & complete linkage

```
Dmat = dist(AuthorData,method="canberra")
com.hc = hclust(Dmat,method="complete")
res.com = cutree(com.hc,4)
table(res.com,TrueAuth)
```

```
##        TrueAuth
## res.com Austen London Milton Shakespeare
##       1    316    219      0         173
##       2      1     74      0           0
##       3      0      3      0           0
##       4      0      0     55           0
```

```
plot(com.hc,cex=.5)
```

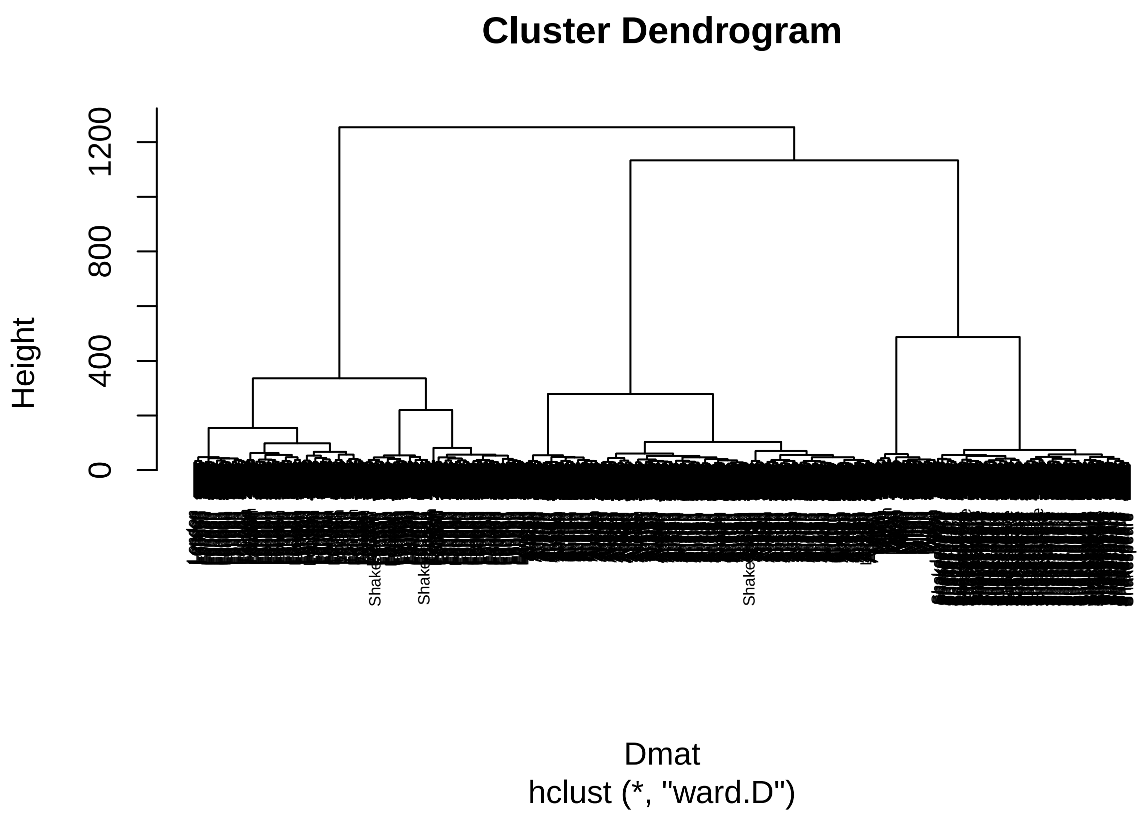


Which linkage is best? Why?
canberra distance & ward.D linkage

```
Dmat = dist(AuthorData,method="canberra")
ward.hc = hclust(Dmat,method="ward.D")
res.ward = cutree(ward.hc,4)
table(res.ward,TrueAuth)
```

```
##           TrueAuth
## res.ward Austen London Milton Shakespeare
##        1    312      1      0           1
##        2      1      3      0         170
##        3      4    292      0           2
##        4      0      0     55           0
```

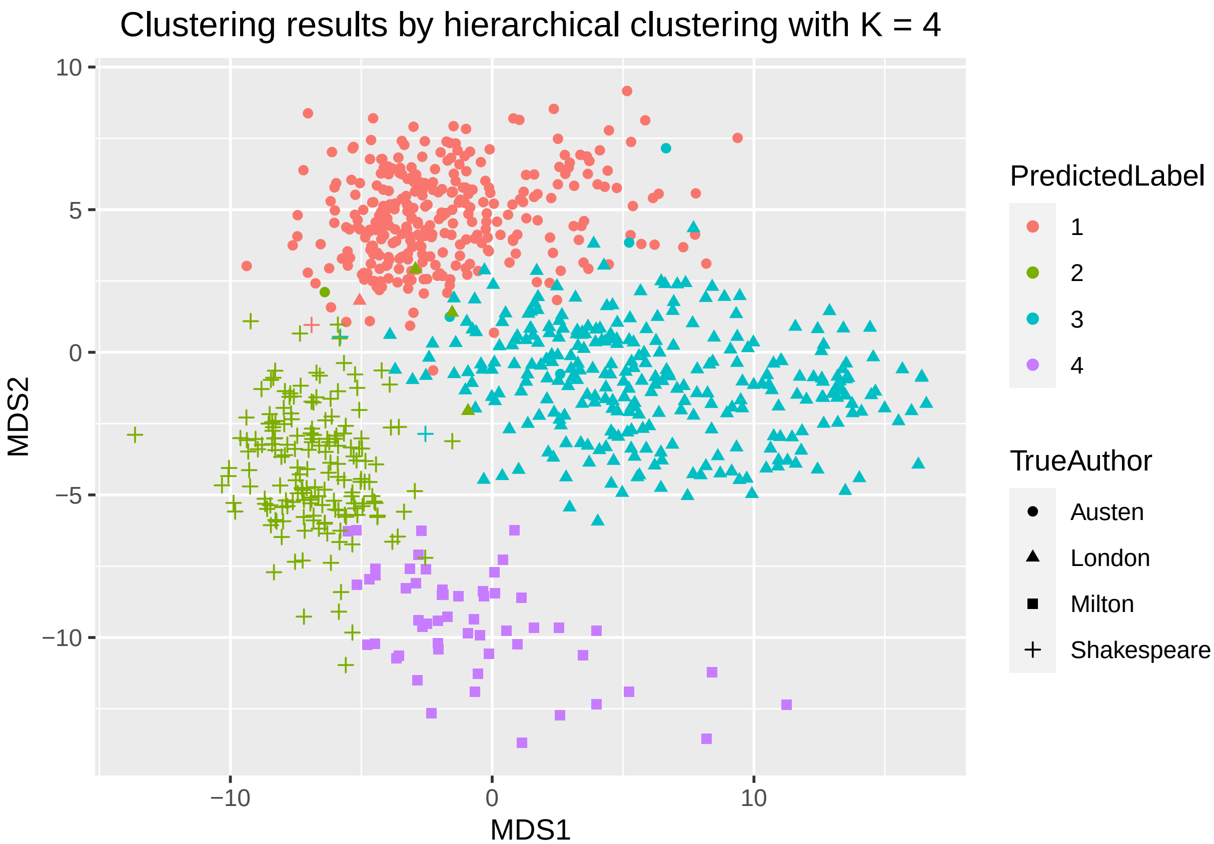
```
plot(ward.hc,cex=.5)
```



We can see that canberra distance and ward.D linkage give excellent clustering results.

Do any preform terribly? Why?
Visualizing hierarchical clustering results using MDS.

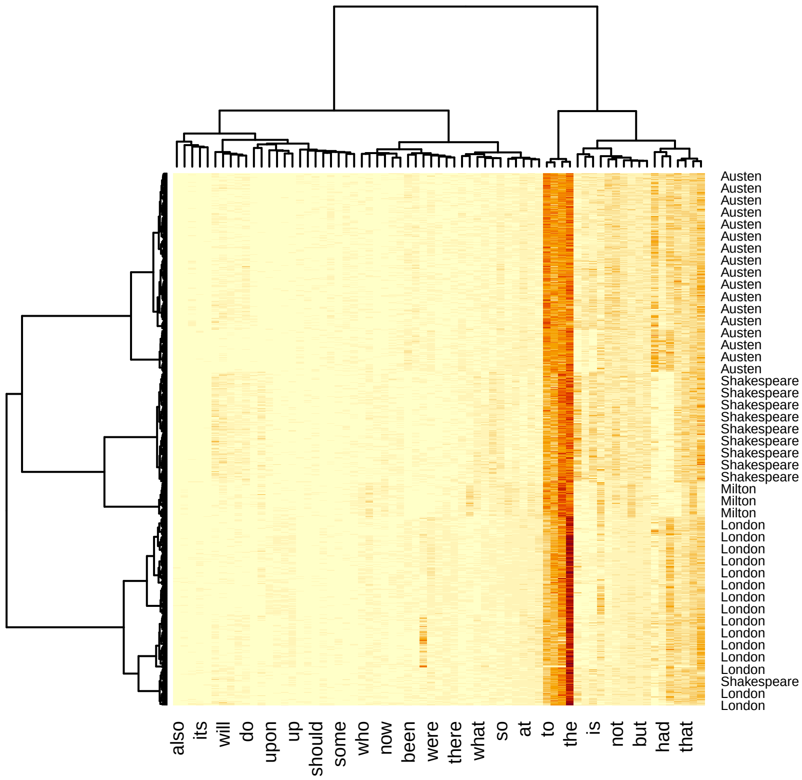
```
PredData = data.frame(cbind(MDSData[,1:2],res.ward,rownames(AuthorData)))
colnames(PredData) = c("MDS1","MDS2","PredictedLabel","TrueAuthor")
PredData$PredictedLabel = factor(PredData$PredictedLabel)
ggplot(PredData) +
  geom_point(mapping=aes(x = MDS1,y= MDS2,color = PredictedLabel,shape= TrueAuthor)) +
  ggtitle("Clustering results by hierarchical clustering with K = 4") +
  theme(plot.title = element_text(hjust = 0.5))
```



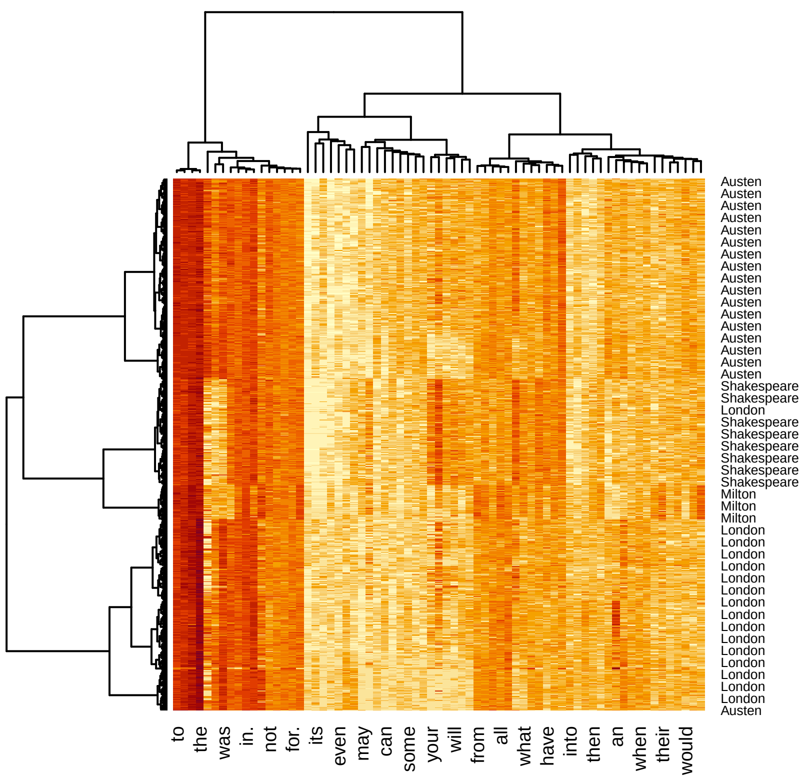
Problem 4 - Biclustering

Cluster heatmap

```
heatmap(AuthorData,
        distfun=function(x)dist(x,method="canberra"),
        hclustfun=function(x)hclust(x,method="ward.D"))
```




```r
heatmap(log(AuthorData+1),
        distfun=function(x)dist(x,method="canberra"),
        hclustfun=function(x)hclust(x,method="ward.D"))
```



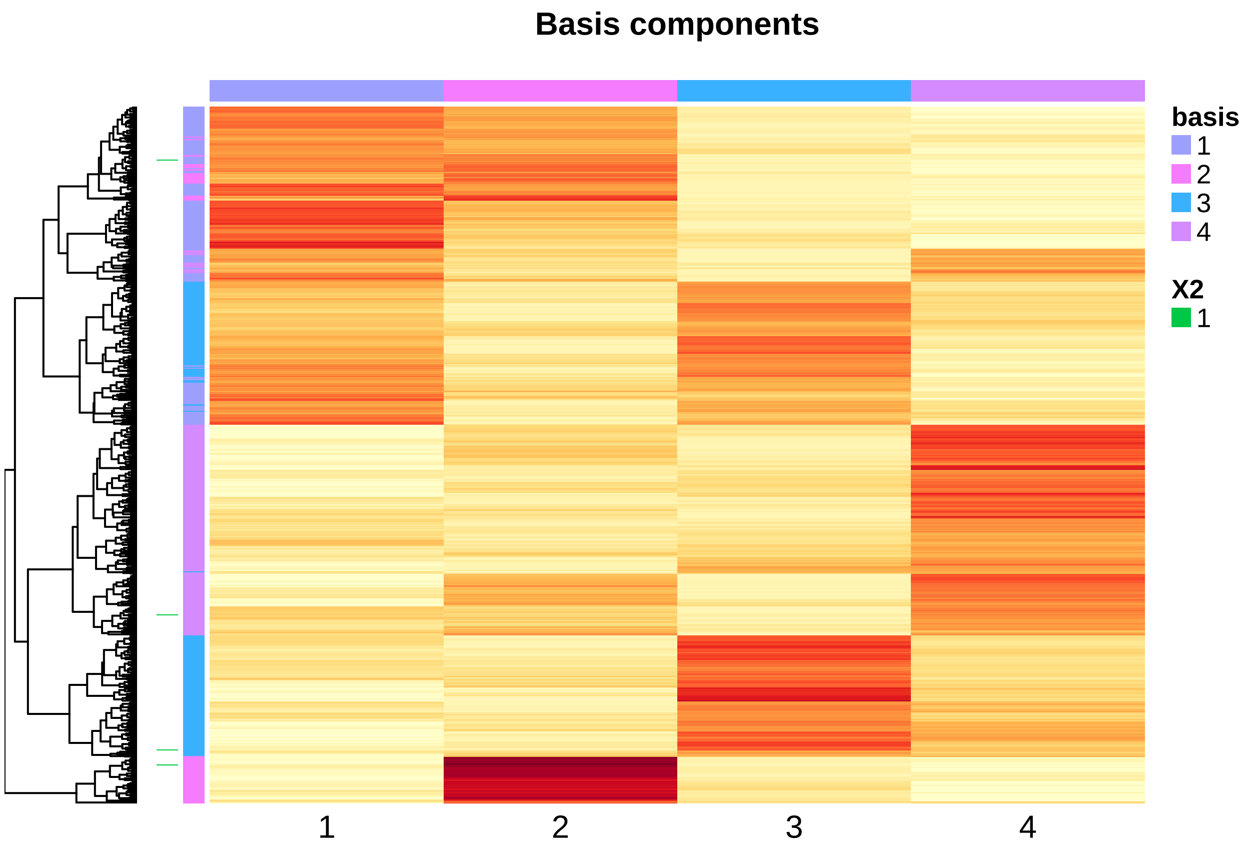
Problem 5 - NMF

```r
K = 4
nmffit = nmf(AuthorData,rank=K)
W = basis(nmffit)
H = coef(nmffit)
```

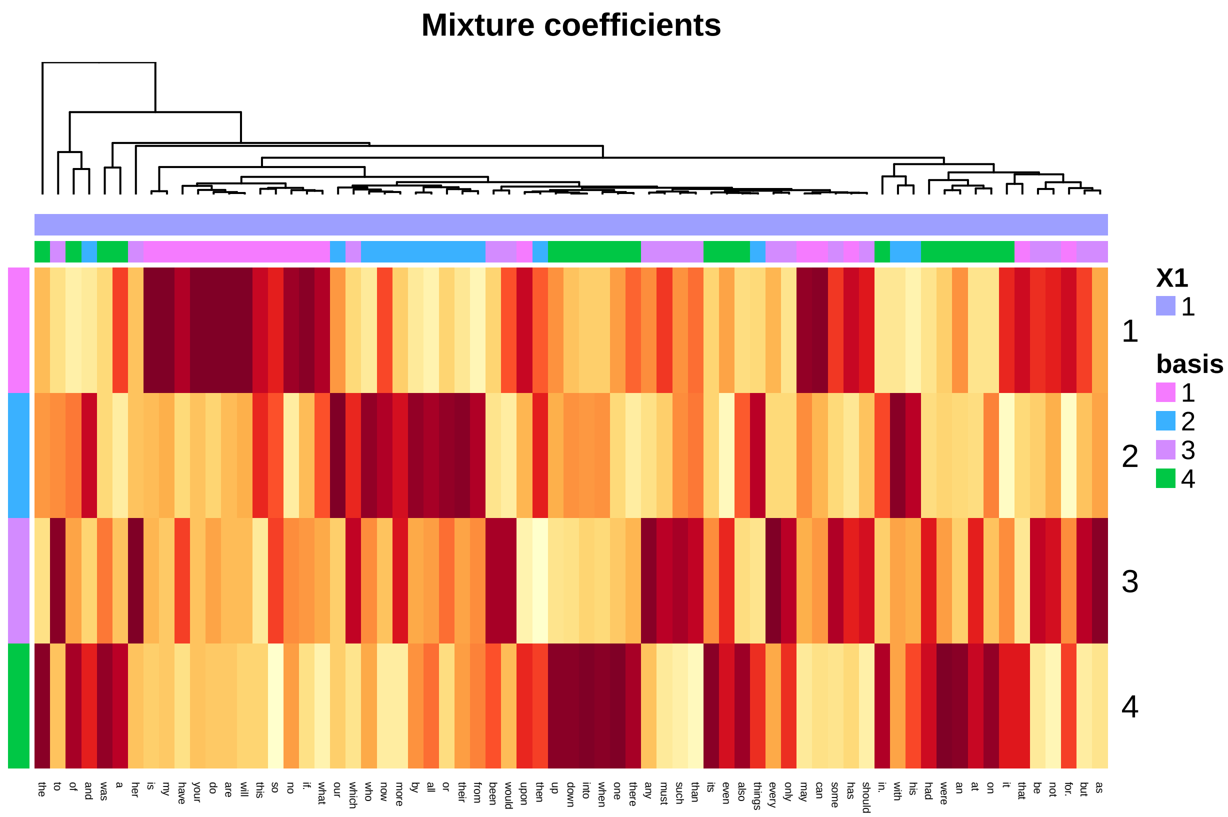
```r
cmap = apply(W,1,which.max)
table(cmap,TrueAuth)
```

```
##     TrueAuth
## cmap Austen London Milton Shakespeare
##    1     51     25      0         143
##    2      0      1     55          30
##    3    263      2      0           0
##    4      3    268      0           0
```

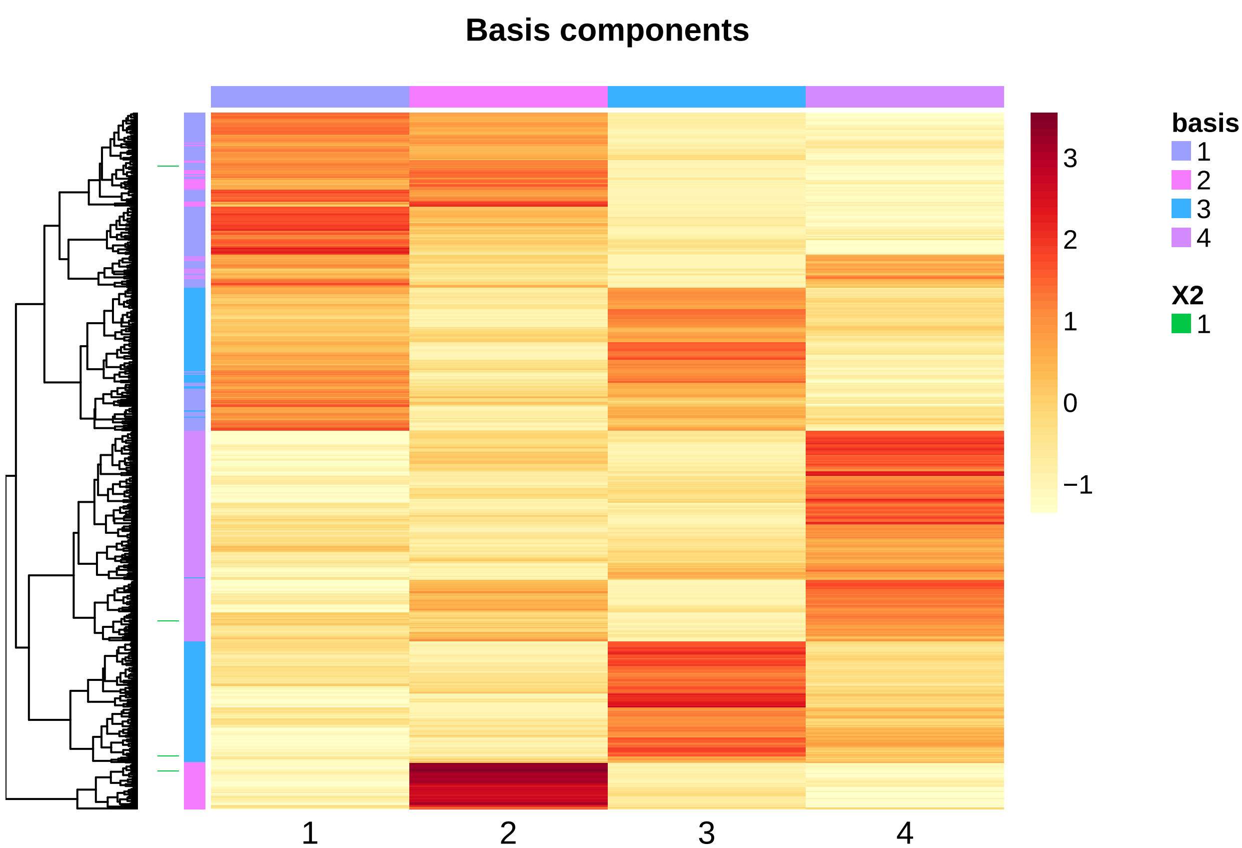
```r
basismap(nmffit,annRow=rownames(AuthorData),scale="col",legend=FALSE)
```



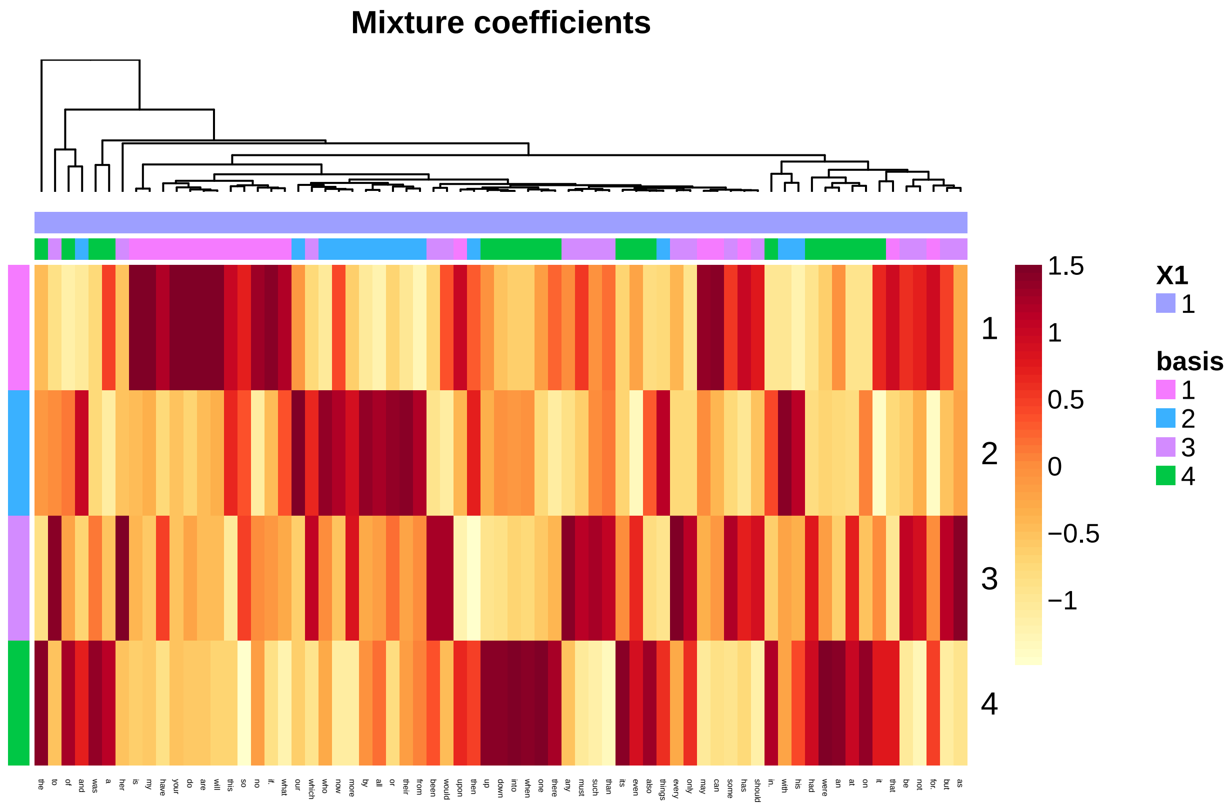

```r
coefmap(nmffit,annCol=colnames(AuthorData),scale="col",legend=FALSE)
```




```
basismap(nmffit,annRow=rownames(AuthorData),scale="col",legend=T)
```



```
coefmap(nmffit,annCol=colnames(AuthorData),scale="col",legend=T)
```



Which words are most important for distinguishing authors?