TEXT-MINING THE BIBLE

Vignesh J Muralidharan

October 6, 2018

```
library(textmineR) ;library(tidyverse) ;library(factoextra)
library(cluster) ; library(NbClust) ;library(fpc) ; library(wordcloud)
library(dendroextras) ; library(dendextend) ;library(mclust)
library(dbscan) ; library(dplyr); library(e1071) ; library(seriation)
library(DT);library(arulesViz);library(arulesCBA); library(dplyr)

bible<-read.csv("https://raw.githubusercontent.com/vigneshjmurali/Statistical-Predictive-Modelling
/master/Datasets/bible_asv.csv")
dim(bible)
## [1] 31103     8

# CREATING FACTOR VARIABLE FOR VARIABLE BOOKS
bible_bt=aggregate(Testaments~Books,data=bible,FUN = unique,collapse="" )
bible_bt$Testaments=as.factor(ifelse(bible_bt$Testaments==bible_bt$Testaments[1],1,2))# Creating l
evels for books as OT =1 & NT =2

levels(bible$Sections)
## [1] "Apostles" "Gospels"  "History"  "Law"      "Paul"     "Prophets"
## [7] "Wisdom"
bible_bs=aggregate(Sections~Books, data=bible, FUN = unique, collapse="")
bible_bs$Sections<-ordered(bible_bs$Sections,levels=c('Apostles','Gospels','History','Law','Paul',
'Prophets','Wisdom'))

# CREATING FACTOR VARIABLE FOR VARIABLE CHAPTERS
bible_cht=aggregate(Testaments~Chapters,data=bible,FUN=unique, collapse="")
bible_cht$Testaments=as.factor(ifelse(bible_cht$Testaments==bible_cht$Testaments[1],1,2))

bible_chs=aggregate(Sections~Chapters,data=bible,FUN=unique,collapse="")
bible_chs$Sections<-ordered(bible_chs$Sections,levels=c('Apostles','Gospels','History','Law','Paul
','Prophets','Wisdom'))

# CREATING FACTOR VARIABLE FOR VARIABLE VERSES
bible_vt=bible[,c('Testaments','Verses')]
bible_vt$Testaments=as.factor(ifelse(bible_vt$Testaments==bible_vt$Testaments[1],1,2))

bible_vs=bible[,c('Sections','Verses')]
bible_vs$Sections<-ordered(bible_vs$Sections,levels=c('Apostles','Gospels','History','Law','Paul',
'Prophets','Wisdom'))

# CREATING FACTOR VARIABLE FOR VARIABLE TESTAMENTS AND TEXT
bible_tt=aggregate(Testaments~text,data=bible,FUN=unique,collapse="")
bible_tt$Testaments=as.factor(ifelse(bible_tt$Testaments==bible_tt$Testaments[1],1,2))# Creating l
evels for books as OT =1& NT =2

# CREATING FACTOR VARIABLE FOR VARIABLE SECTIONS AND TEXT
bible_st=aggregate(Sections~text,data=bible,FUN=unique,collapse="")
```

Collapsing the text of all the verses into the same books and then the same chapters together before performing clustering analsysis

```
#Collpase text into the same 66 books
attach(bible)
text.Book=c()
for (i in 1:66){
  text.Book[i]=paste(text[Books==as.character(unique(Books)[i])],collapse="")
}
#Collpase text into the same 1189 Chapters
```

```r
text.Chapters=c()
for (i in 1:1189){
  text.Chapters[i]=paste(text[Chapters==as.character(unique(Chapters)[i])],collapse = "")
}
#View(text.Testaments)
#bible_testaments=data.frame(Testaments=unique(Testaments),text=text.Testaments)
bible_books=data.frame(Books=unique(Books),text=text.Book)
bible_chapters=data.frame(Chapters=unique(Chapters),text=text.Chapters)
bible_verses=bible
dim(bible_books);dim(bible_chapters);dim(bible_verses)
## [1] 66  2
## [1] 1189    2
## [1] 31103     8
```

Performing standard text transformations - moving all case to lower, removing numbers, removing punctutation, removing common stopwords, strip whitespace and getting rid of special characters. we will consider n-grams, co-ocurrances, stemming and term document matrix.

```r
my_stopwords1 = c("a", "about", "above", "across", "after", "afterwards", "again", "against", "all
", "almost", "alone", "along", "already", "also","although","always","am","among", "amongst", "amo
ungst", "amount", "an", "and", "another", "any","anyhow","anyone","anything","anyway", "anywhere",
"are", "around", "as", "at", "back","be","became", "because","become","becomes", "becoming", "been
", "before", "beforehand", "behind", "being", "below", "beside", "besides", "between", "beyond", "
bill", "both", "bottom","but", "by", "call", "can", "cannot", "cant", "co", "con", "could", "could
nt", "cry", "de", "describe", "detail", "do", "done", "down", "due", "during", "each", "eg", "eigh
t", "either", "eleven","else", "elsewhere", "empty", "enough", "etc", "even", "ever", "every", "ev
eryone", "everything", "everywhere", "except", "few", "fifteen", "fify", "fill", "find", "fire", "
first", "five", "for", "former", "formerly", "forty", "found", "four", "from", "front", "full", "f
urther", "get", "give", "go", "had", "has", "hasnt", "have", "he", "hence", "her", "here", "hereaf
ter", "hereby", "herein", "hereupon", "hers", "herself", "him", "himself", "his", "how", "however"
, "hundred", "ie", "if", "in", "inc", "indeed", "interest", "into", "is", "it", "its", "itself", "
keep", "last", "latter", "latterly", "least", "less", "ltd", "made", "many", "may", "me", "meanwhi
le", "might", "mill", "mine", "more", "moreover", "most", "mostly", "move", "much", "must", "my",
"myself", "name", "namely", "neither", "never", "nevertheless", "next", "nine", "no", "nobody", "n
one", "noone", "nor", "not", "nothing", "now", "nowhere", "of", "off", "often", "on", "once", "one
", "only", "onto", "or", "other", "others", "otherwise", "our", "ours", "ourselves", "out", "over"
, "own","part", "per", "perhaps", "please", "put", "rather", "re", "same", "see", "seem", "seemed"
, "seeming", "seems", "serious", "several", "she", "should", "show", "side", "since", "sincere", "
six", "sixty", "so", "some", "somehow", "someone", "something", "sometime", "sometimes", "somewher
e", "still", "such", "system", "take", "ten", "than", "that", "the", "their", "them", "themselves"
, "then", "thence", "there", "thereafter", "thereby", "therefore", "therein", "thereupon", "these"
, "they", "thickv", "thin", "third", "this", "those", "though", "three", "through", "throughout",
"thru", "thus", "to", "together", "too", "top", "toward", "towards", "twelve", "twenty", "two", "u
n", "under", "until", "up", "upon", "us", "very", "via", "was", "we", "well", "were", "what", "wha
tever", "when", "whence", "whenever", "where", "whereafter", "whereas", "whereby", "wherein", "whe
reupon", "wherever", "whether", "which", "while", "whither", "who", "whoever", "whole", "whom", "w
hose", "why", "will", "with", "within", "without", "would", "yet", "you", "your", "yours", "yourse
lf", "yourselves", "the")
my_stopwords2 = c('thou','thee','thy','ye','shall','shalt','lo','unto','hath','thereof','hast', 's
et','thine','art','yea','midst','wherefore','wilt','thyself')

#Canonical Groupings of the Bible
Testaments=c(rep('OT',39),rep('NT',27))
Sections=c(rep('Law',5),rep('History',12),rep('Wisdom',5),rep('Prophets',17),rep('Gospels',5),rep(
'Paul',13),rep("Apostles",9))
bible_new =data.frame(Books=unique(Books),Testaments=as.factor(c(rep("OT",39),rep("NT",27))),
Sections=as.factor(c(rep("Law",5),rep("History",12),rep("Wisdom",5),rep("Prophets",17),rep("Gospel
s",5),rep("Paul",13),rep("Apostles",9))),
                     text=text.Book)
```

CLUSTERING ON THE TEXT OF 66 BOOKS

Turning the sentences to document term matrix (DTM)

```r
dtm_b <- CreateDtm(bible_books$text,doc_names = bible_books$Books,ngram_window = c(1, 7),
                   stopword_vec = c(tm::stopwords("english"),tm::stopwords("SMART"),
```

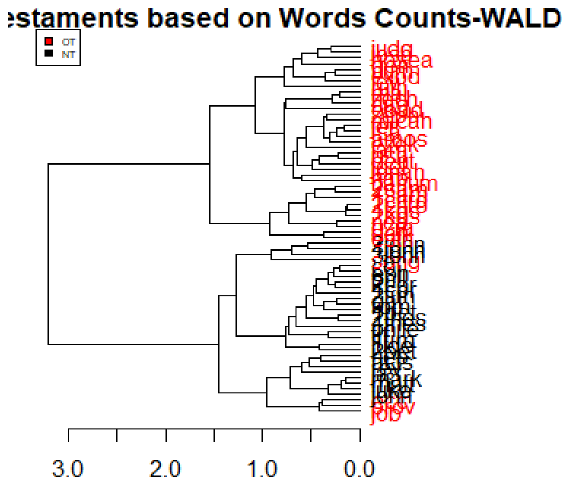
```
                                              my_stopwords1, my_stopwords2),
#stem_lemma_function = function(x) SnowballC::wordStem(x, "porter"),
lower = TRUE, remove_punctuation = TRUE, remove_numbers = FALSE)
##
  |=================================                              |  53%
  |
  |========================================                       |  64%
  |
  |===============================================                |  74%
  |
  |======================================================         |  85%
  |
  |=============================================================  |  95%
  |
  |================================================================| 100%
##
  |
  |===========================                                    |  42%
  |
  |=================================                              |  53%
  |
  |========================================                       |  64%
  |
  |===============================================                |  74%
  |
  |======================================================         |  85%
  |
  |=============================================================  |  95%
  |
  |================================================================| 100%
# explore basic frequencies & acurate vocabulary
tf <- TermDocFreq(dtm_b)
# Keep only words appearing more than 2 times, AND in more than 1 document
vocabulary <- tf$term[tf$term_freq>2 & tf$doc_freq>1]
dtm_b <- dtm_b[ , vocabulary]
## Use term raw Frequency counts
## Calculating document-to-document COSINE SIMILARITY (scalar product)
csim_b <- dtm_b / sqrt(rowSums(dtm_b*dtm_b))
csim_b <- csim_b %*% t(csim_b)

# Turn that cosine similarity matrix into a distance matrix
dist.mtx_b <- 1-csim_b
# Calc Hellinger Dist (x = mymat)
# dist.mtx=CalcHellingerDist(as.matrix(dtm))
#Canonical Groupings of the Bible
Testaments=c(rep('OT',39),rep('NT',27))
Sections=c(rep('Law',5), rep('History',12),rep('Wisdom',5),rep('Prophets',17),
rep('Gospels',5),rep('Paul',13),rep("Apostles",9))
```

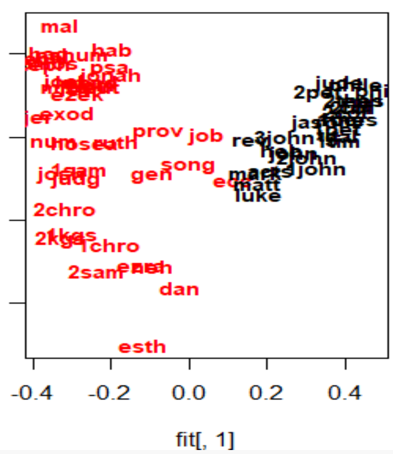
Dendrograms

Dendrogram for the 2 Testaments on the text f the 66 books & Dendrogram for the 7 levels of Sections in the Bible

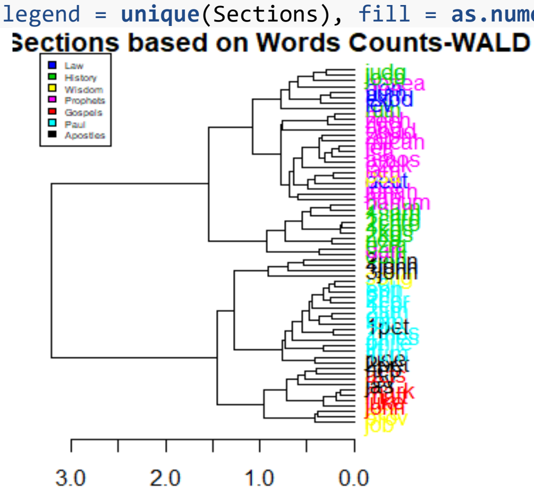
```
#Dendrogram for the 2 Testaments on the text f the 66 books
# Using the term raw frequency counts with dendrograms using wald linkage
hc.wald=hclust(as.dist(dist.mtx_b),'ward.D2')
dend=as.dendrogram(hc.wald)
#Coloring the leaves according to 'Testaments'
labels_colors(dend)<-as.numeric(as.factor(Testaments[hc.wald$order]))
#Change labels font size
dend<-set(dend,"labels_cex",1.0)
plot(dend,horiz = TRUE,main='Dend of 2 Testaments based on Words Counts-WALD')
legend("topleft", cex=0.45, legend = unique(Testaments), fill = as.numeric(as.factor(unique(Testam
ents))))
```



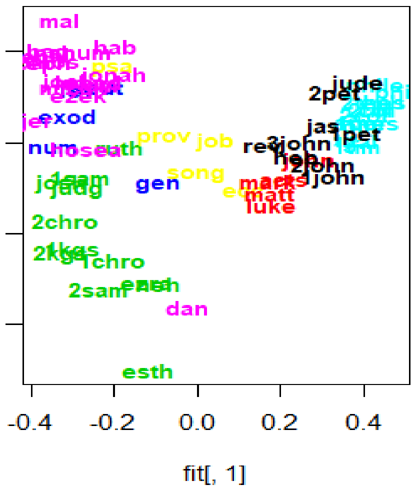

```
#MDS Plot
fit<-cmdscale(as.dist(dist.mtx_b),k=2)
#Two Testaments
plot(fit[,2]~fit[,1],type='n')
text(x = fit[,1], y = fit[,2], labels = row.names(fit), col=unclass(as.factor(Testaments)), cex=.95, font=2)
mtext( cex = 1, text = "Two Testaments of the Bible based on Words Counts", line=2,outer=FALSE)
```




```
#Dendrogram for the 7 Levels of Sections in the Bible
hc.wald=hclust(as.dist(dist.mtx_b),'ward.D2');dend=as.dendrogram(hc.wald)
#Coloring the leaves according to 'Sections'
labels_colors(dend)<-as.numeric(as.factor(Sections[hc.wald$order]))
#Change labels font size
dend<-set(dend,"labels_cex",1.0) ; plot(dend,horiz=TRUE,main='Dend of 7 Sections based on Words Counts-WALD')
legend("topleft", cex=0.45, legend = unique(Sections), fill = as.numeric(as.factor(unique(Sections))))
```



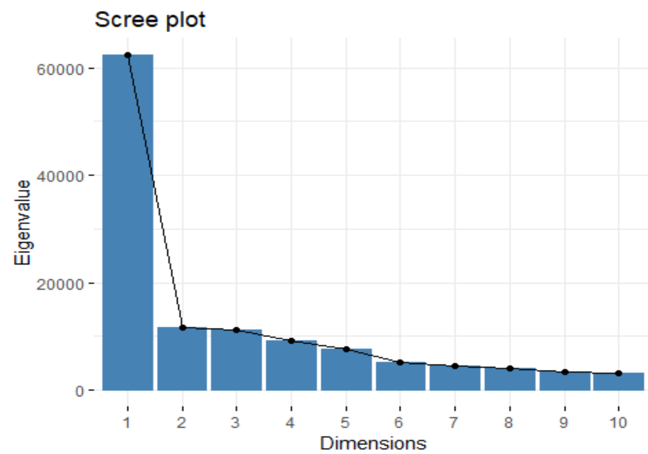
```r
#MDS Plot
fit<-cmdscale(as.dist(dist.mtx_b),k=2)
#Seven Sections
plot(fit[,2]~fit[,1],type='n')
text(x = fit[,1], y = fit[,2], labels = row.names(fit), col=unclass(as.factor(Sections)), cex=.95, font=2)
mtext( cex = 1, text = "Seven Sections of the Bible based on Words Counts", line=2,outer=FALSE)
```



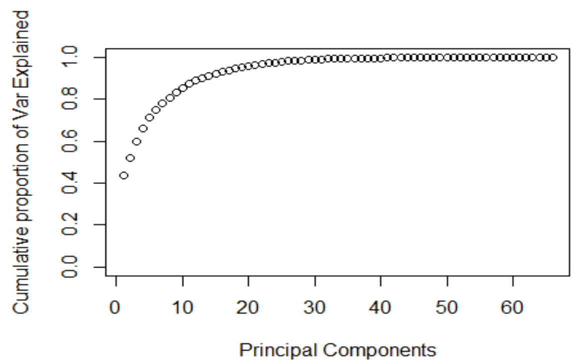
For Testaments - Both the Dendrogram and the MSD plot shows that the resulting 2 clusters match the 2 Testaments pretty well For Sections - Both the Dendrogram and the MSD plot shows that the resulting 7 clusters match the 7 Sections well. Here its better to find some misclassifications.

PRINCIPLE COMPONENT ANALYSIS

```r
#Transforming the dtm into a matrix
m_b<-as.matrix(dtm_b);dtm_b.pca=prcomp(m_b) #did not scale the data
#Here the rotations measure provides the principal component loadings.
#Each column of the rotation matrix contains the principal component loading vector
dtm_b.pca$rotation[1:5,1:5] #Showing firt 5 PC and first 5 rows
##                           PC1           PC2           PC3           PC4
## round_cut       -1.648058e-04 -0.0008648294 -0.0006715227 -0.0001366474
## jehovah_god_die -4.146433e-04  0.0001636758  0.0004045446 -0.0003189592
## worthy_unloose   2.809317e-05  0.0002548109  0.0002151326  0.0009181075
## saul_meet       -1.718941e-04  0.0009938088 -0.0004880269  0.0002247653
## season_jesus     2.616924e-05  0.0003256117  0.0001408742  0.0008984095
##                           PC5
## round_cut       -7.600066e-05
## jehovah_god_die  2.579996e-04
## worthy_unloose  -8.116030e-04
## saul_meet        6.699454e-04
## season_jesus    -9.047611e-04
dim(dtm_b.pca$x)
## [1] 66 66
#Standard deviation of each principal component
dtm_b.sd=dtm_b.pca$sdev; dtm_b.var=dtm_b.pca$sdev^2 #To compute variance
dtm_b.var[1:5]
## [1] 62365.711 11599.182 11129.915  9067.942  7667.156
#Proportion of Variance explained
pve=dtm_b.var/sum(dtm_b.var) ; cumsum(pve[1:10])
##  [1] 0.4374061 0.5187578 0.5968182 0.6604168 0.7141909 0.7504946 0.7811174
##  [8] 0.8091771 0.8327983 0.8554135
#Plotting the principal components with proportion of variance explained
#plot(dtm_b.pca,type="lines")
fviz_screeplot(dtm_b.pca,np=10,choice="eigenvalue")
```




```
plot(cumsum(pve),xlab="Principal Components", ylab="Cumulative proportion of Var Explained", ylim=c(0,1),type='b')
```

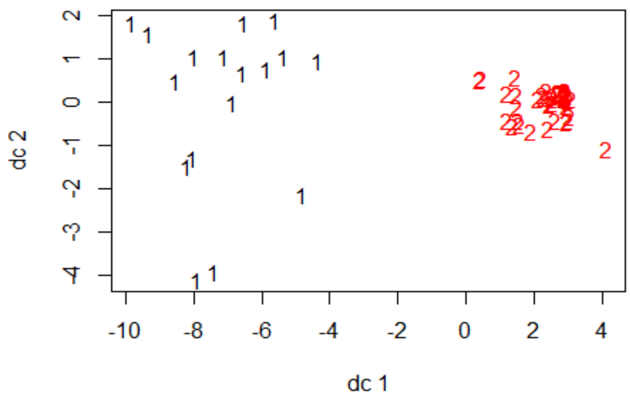



```
which.max(cumsum(pve)[cumsum(pve)<0.90])
## [1] 12
#From this we can see that within 12 PC we can cover almost 90% of the variance of the data
dtm_bnew=as.data.frame(dtm_b.pca$x[,1:12]); dtm_bnew1=dtm_b.pca$x[,1:12]
```

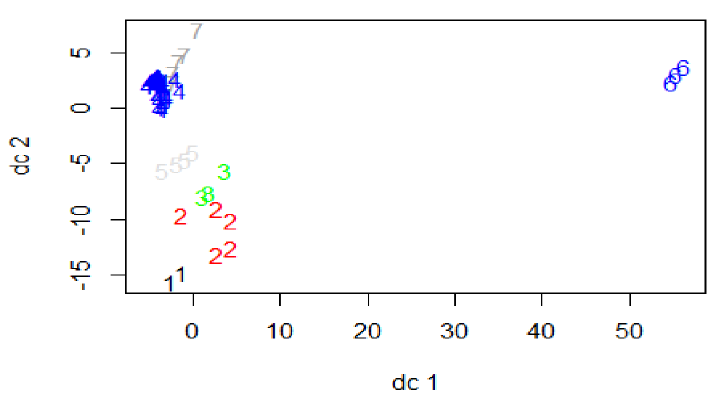
Therefore the first 12 PC were chosn as the new variables

K-Means Clustering

```
# Performing K-means clustering with k=2
set.seed(2)
km_2.fit=kmeans(dtm_bnew,2,nstart=50); attributes(km_2.fit)
## $names
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
## $class
## [1] "kmeans"
# Both the Testaments "OT" and "NT" is labeled as '1' & '2'
y_k2=table(km_2.fit$cluster, bible_bt$Testaments) ; y_k2
##       1  2
##    1  9  8
##    2 30 19
#Accuracy
mean(km_2.fit$cluster==bible_bt$Testaments)
## [1] 0.4242424
#Misclassification rate
misrate_k2<-1-sum(diag(y_k2))/sum(y_k2) ; misrate_k2
## [1] 0.5757576
#Centroid plot with against 1st and 2nd discriminant functions
plotcluster(dtm_bnew,km_2.fit$cluster)
```



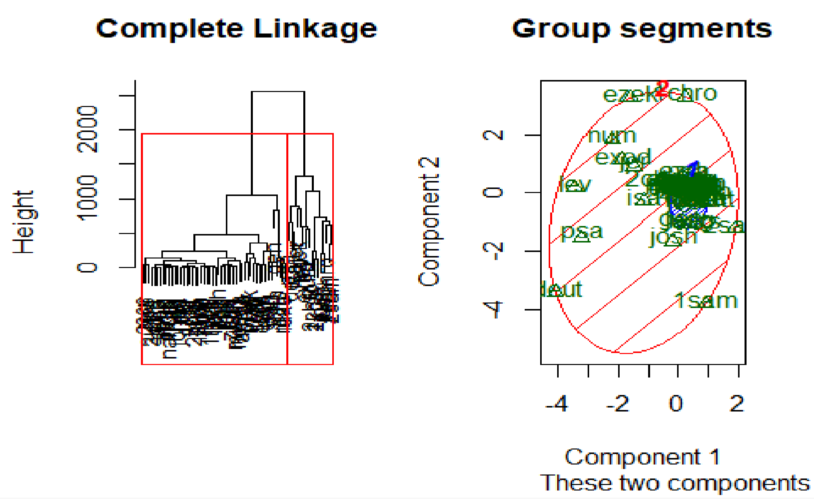

```
# Performing K-means clustering with k=7
set.seed(4); km_7.fit=kmeans(dtm_bnew,7,nstart = 50) ; attributes(km_7.fit)
## $names
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
## $class
## [1] "kmeans"
#7 Sections('Apostles'-'1', 'Gospels'-'2',  'History'-'3', 'Law'-'4', 'Paul'-'5','Prophets'-'6','W
isdom'-'7') were labeled
y_k7=table(km_7.fit$cluster,bible_bs$Sections) ; y_k7
##     Apostles Gospels History Law Paul Prophets Wisdom
##   1        1       0       0   0    0        1      0
##   2        1       0       3   0    1        0      0
##   3        0       0       0   1    1        0      1
##   4        4       4       4   4    9       15      4
##   5        1       0       2   0    1        0      0
##   6        1       0       1   0    1        0      0
##   7        1       1       2   0    0        1      0
mean(km_7.fit$cluster == bible_bs$Sections)
## [1] 0
misrate_k7<-1-sum(diag(y_k7))/sum(y_k7)  ; misrate_k7
## [1] 0.9090909
# Centroid Plot against 1st 2 discriminant functions
plotcluster(dtm_bnew, km_7.fit$cluster)
```



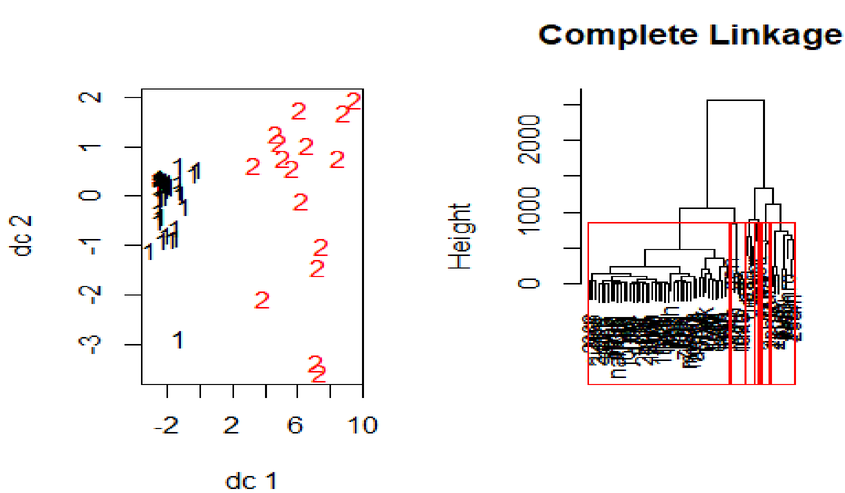
The K-means clustering we can see that the missclassification rate on the 2 testaments and 7 sections are high. But this problem can also be due to the set.seed . When i change the set.seed the missclassification is getting decreased for k=2 but not decreasing with k=7

HIERARCHIAL CLUSTERING AFTER PCA

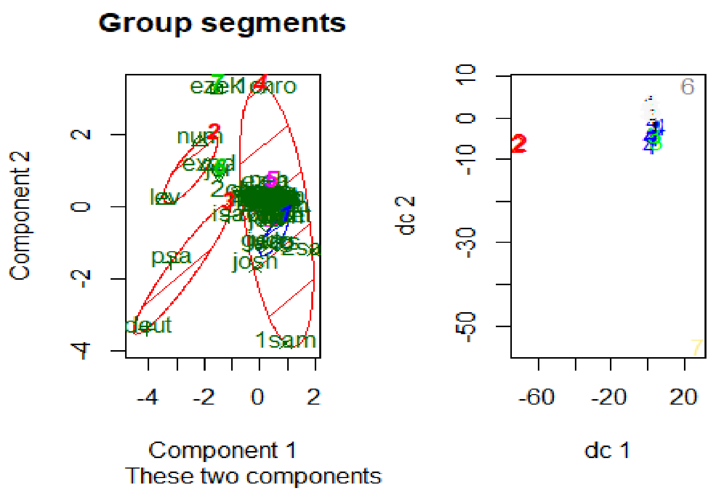
```
par(mfrow=c(1,2))
## Hierarchical Clustering for  k=2 testaments
hc.ward=hclust(dist(dtm_bnew, method = "euclidean"), method="ward.D2")
plot(hc.ward,main="Complete Linkage", xlab="", sub="", cex=.9) #dendrogram
# draw dendogram with red borders around the 2 clusters
rect.hclust(hc.ward,k=2,border="red")
groups2=cutree(hc.ward,2)# cut tree into 5 clusters
#Accuracy and misclassification rate
y_h2<-table(groups2,bible_bt$Testaments) ;y_h2
## groups2  1  2
##       1 31 19
##       2  8  8
mean(groups2 ==bible_bt$Testaments)
## [1] 0.5909091
misrate_h2<-1-sum(diag(y_h2))/sum(y_h2) ; misrate_h2
## [1] 0.4090909
# 2D representation of the Segmentation:
clusplot(dtm_bnew, groups2, color=TRUE, shade=TRUE,labels=2, lines=0, main= 'Group segments')
```




```
# Centroid Plot against 1st 2 discriminant functions
plotcluster(dtm_bnew, groups2)
## Hierarchical Clustering for  k=7 sections
#dendrogram
plot(hc.ward,main="Complete Linkage", xlab="", sub="", cex=.9)
# draw dendogram with red borders around the 2 clusters
rect.hclust(hc.ward,k=7,border="red")
```



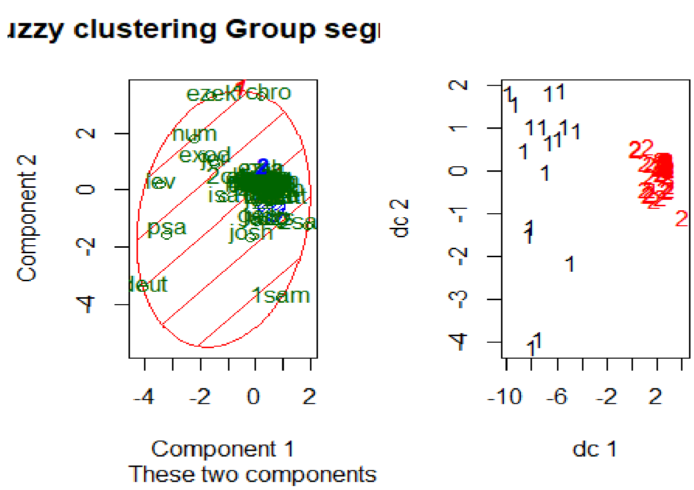
```
groups7=cutree(hc.ward,7)# cut tree into 5 clusters
#Accuracy and misclassification rate
y_h7<-table(groups7,bible_bs$Sections) ;y_h7
## groups7 Apostles Gospels History Law Paul Prophets Wisdom
##       1        1       1       3   0    0        0      0
##       2        1       0       1   0    1        0      0
##       3        1       0       0   0    0        1      1
##       4        2       0       4   0    2        0      0
##       5        4       4       4   4    9       16      4
##       6        0       0       0   0    1        0      0
##       7        0       0       0   1    0        0      0
mean(groups7 ==bible_bs$Sections)
## [1] 0
misrate_h7<-1-sum(diag(y_h7))/sum(y_h7)  ; misrate_h7
## [1] 0.8484848
# 2D representation of the Segmentation:
clusplot(dtm_bnew, groups7, color=TRUE, shade=TRUE,labels=2, lines=0, main= 'Group segments')
# Centroid Plot against 1st 2 discriminant functions
plotcluster(dtm_bnew, groups7)
```



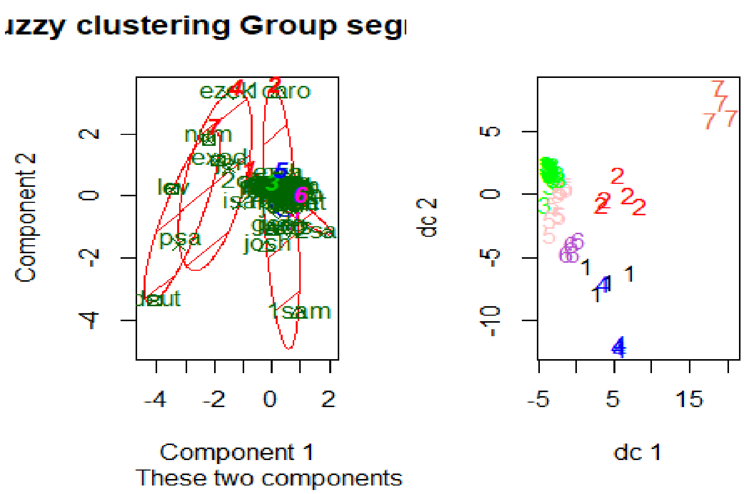
After Hierarchial clustering on the data of the PCA , missclassifiaction rate on the both 2 Testaments and 7 Sections are more high like K-means clustering

FUZZY CLUSTERING

```
par(mfrow=c(1,2))
#k=2 testaments
fuz2 <- cmeans(dtm_bnew, 2, 100, m=2, method="cmeans")
# 2D representation of the Segmentation:
clusplot(dtm_bnew,fuz2$cluster,color=TRUE,shade=TRUE,labels=2,lines=0, main='Fuzzyclustering Group segments')
# Centroid Plot against 1st 2 discriminant functions
plotcluster(dtm_bnew, fuz2$cluster)
```



```
#Accuracy and misclassification rate
y_f2<-table(fuz2$cluster,bible_bt$Testaments) ; y_f2
##       1  2
##   1  9  8
##   2 30 19
mean(fuz2$cluster ==bible_bt$Testaments)
## [1] 0.4242424
misrate_f2<-1-sum(diag(y_f2))/sum(y_f2) ; misrate_f2
## [1] 0.5757576
#k=7 sections
fuz7 <- cmeans(dtm_bnew, 7, 100, m=2, method="cmeans")
# 2D representation of the Segmentation:
clusplot(dtm_bnew,fuz7$cluster,color=TRUE,shade=TRUE,labels=2,lines=0,main='Fuzzy clustering Group segments')
# Centroid Plot against 1st 2 discriminant functions
plotcluster(dtm_bnew, fuz7$cluster)
```



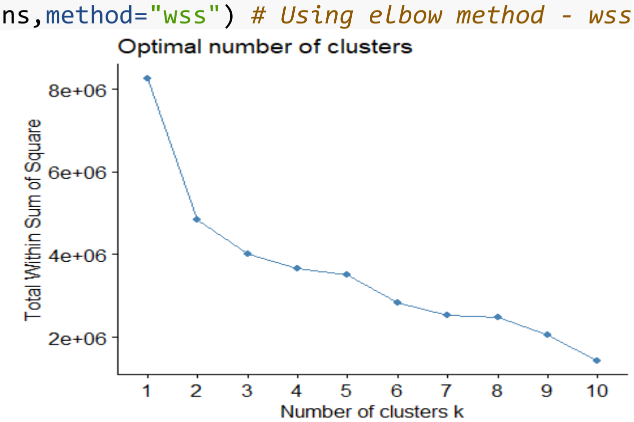

```
#Accuracy and misclassification rate
y_f7<-table(fuz7$cluster,bible_bs$Sections) ;y_f7
##     Apostles Gospels History Law Paul Prophets Wisdom
##   1        1       0       2   0    1        0      0
##   2        1       0       3   0    1        0      0
##   3        4       3       4   3    6       12      3
##   4        0       0       0   1    1        1      1
##   5        0       1       0   1    3        4      1
##   6        1       1       2   0    0        0      0
##   7        2       0       1   0    1        0      0
mean(fuz7$cluster ==bible_bs$Sections)
## [1] 0
misrate_f7<-1-sum(diag(y_f7))/sum(y_f7) ; misrate_f7
## [1] 0.8636364
```

The missclassification rate on the 2 Testaments and 7 Sections are high after doing Fuzzy Clustering.

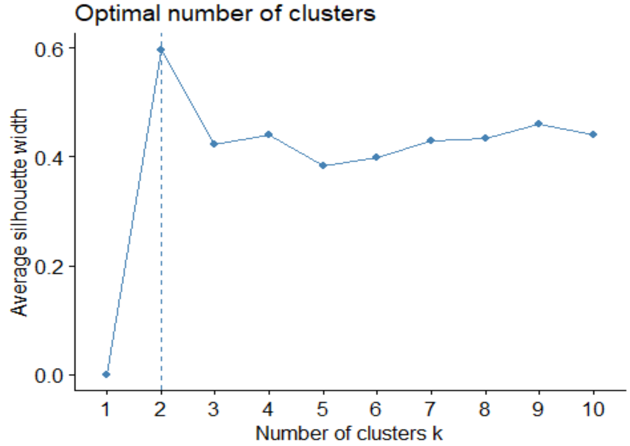
NB-CLUST

NbClust proposes that 2 is the best clustering method for this new dataset

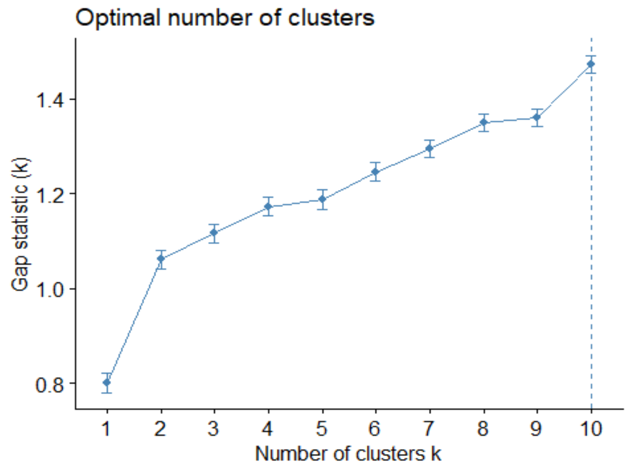
```
fviz_nbclust(dtm_bnew1,kmeans,method="wss") # Using elbow method - wss
```



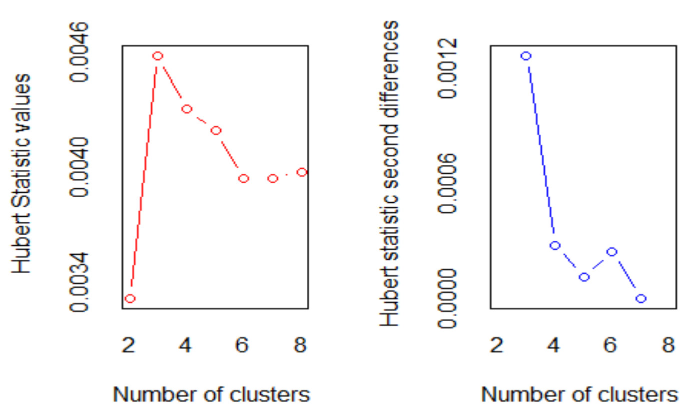
```r
fviz_nbclust(dtm_bnew1,kmeans,method="silhouette") #Using silhouette method
```



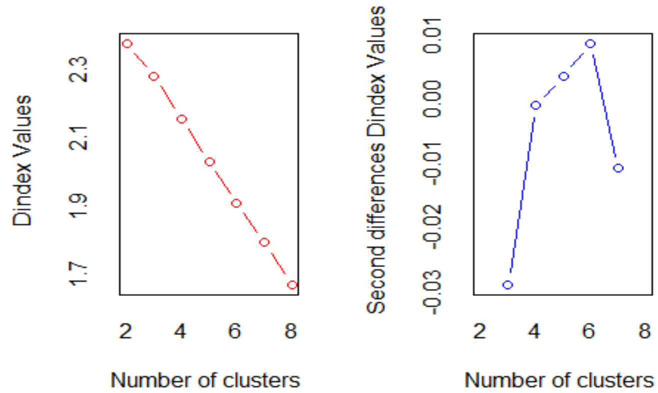

```r
fviz_nbclust(dtm_bnew1,kmeans,method="gap_stat") #Using gap_stat method
```




```r
mito.nbclust<-dtm_bnew1 %>% #Using NbClust
  scale() %>%
  NbClust(distance="euclidean",min.nc=2,max.nc=8,method="complete",index="all")
```



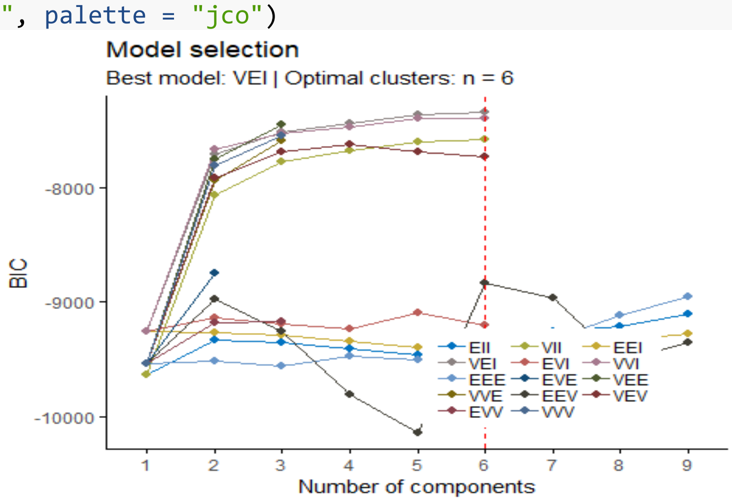

```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##           significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
```



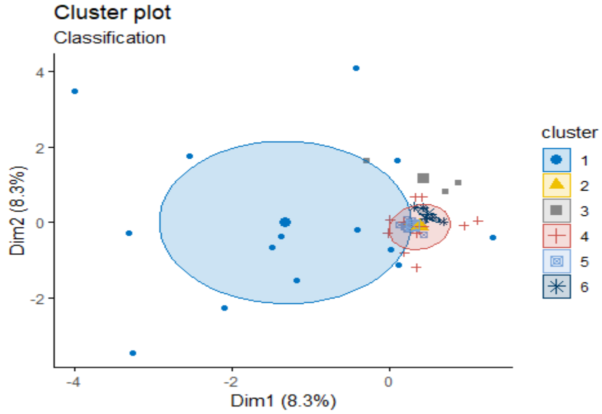

```
## *** : The D index is a graphical method of determining the number of clusters.
##                In the plot of D index, we seek a significant knee (the significant peak in Dindex
## second differences plot) that corresponds to a significant increase of the value of measure
## *******************************************************************
## * Among all indices:
## * 9 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 8 proposed 8 as the best number of clusters
##                    ***** Conclusion *****
## * According to the majority rule, the best number of clusters is  2
## *******************************************************************
```

MODEL BASED CLUSTERING (MDS)

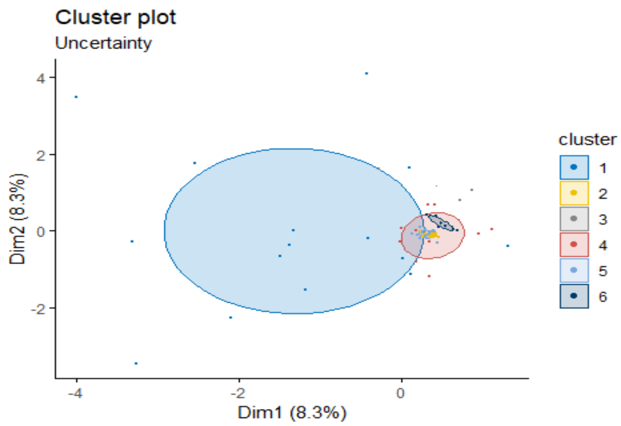
```
mb.fit <- Mclust(dtm_bnew)
summary(mb.fit) # display the best model
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
## Mclust VEI (diagonal, equal shape) model with 6 components:
##  log.likelihood  n df       BIC       ICL
##       -3479.037 66 94 -7351.902 -7352.688
## Clustering table:
##  1  2  3  4  5  6
## 14 19  3 10 15  5
mb.fit$modelName # Optimal selected model ==> "VVV"
## [1] "VEI"
mb.fit$G  # Optimal number of cluster => 6
## [1] 6
# BIC values used for choosing the number of clusters
fviz_mclust(mb.fit, "BIC", palette = "jco")
```



```r
# Classification: plot showing the clustering
fviz_mclust(mb.fit, "classification", geom = "point", pointsize = 1.5, palette = "jco")
## Too few points to calculate an ellipse
```



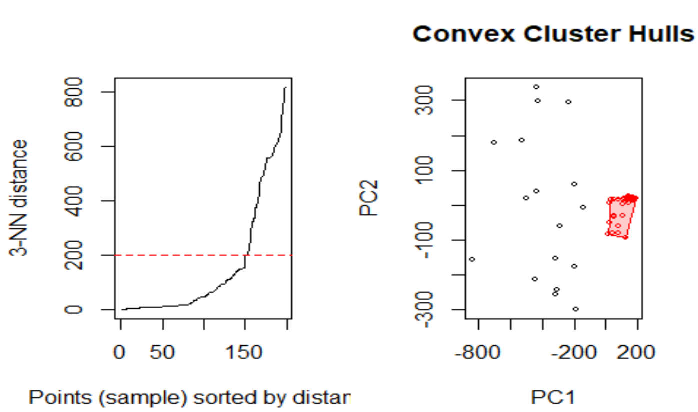

```r
# Classification uncertainty
fviz_mclust(mb.fit, "uncertainty", palette = "jco")## Too few points to calculate an ellipse
```



Results from Model based clustering shows that the optimal number of clusters is 6 and not 2 or 7 which we found from the kmeans, fuzzy, hirachal and nbclust

DENSITY BASED CLUSTERING

```r
par(mfrow=c(1,2)) ;set.seed(123)
# determining the optimal eps value
dbscan::kNNdistplot(dtm_bnew, k =  3); abline(h = 200, lty = 2,col="red")
dbm <- fpc::dbscan(dtm_bnew, eps = 200,  MinPts = 5) ;dbm
## dbscan Pts=66 MinPts=5 eps=200
##         0  1
## border 17  0
## seed    0 49
## total  17 49
#Display the hull plot
hullplot(dtm_bnew, dbm$cluster)
```



MISSCLASSIFICATON RATE OF THE TESTAMENTS AND THE SECTIONS

```
# missclassification rate on 2 Testaments
cv_error_rate2 <- rbind(misrate_k2,misrate_h2,misrate_f2)
rownames(cv_error_rate2) <- (c('Kmeans Clustering','Hierarchical Clustering','Fuzzy Clustering'))
colnames(cv_error_rate2) <- 'cv_error_rate2' ; round(cv_error_rate2, 4)
##                           cv_error_rate2
## Kmeans Clustering                 0.5758
## Hierarchical Clustering           0.4091
## Fuzzy Clustering                  0.5758
# missclassification rate on 7 Sections
cv_error_rate7 <- rbind(misrate_k7,misrate_h7, misrate_f7)
rownames(cv_error_rate7) <- (c('Kmeans Clustering', 'Hierarchical Clustering','Fuzzy Clustering'))
colnames(cv_error_rate7) <- 'cv_error_rate' ; round(cv_error_rate7, 4)
##                           cv_error_rate
## Kmeans Clustering                0.9091
## Hierarchical Clustering          0.8485
## Fuzzy Clustering                 0.8636
```

Clustering Groups to tabulate the groups of clusters

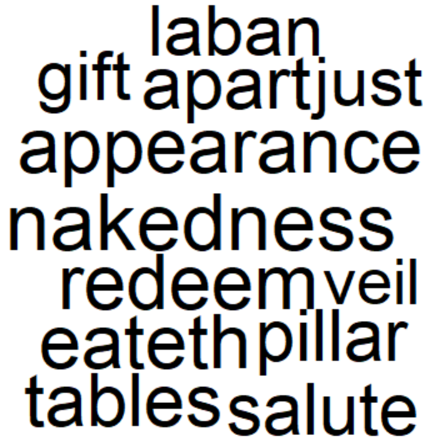
```
bible.group_sections<-data.frame(dtm_bnew,km_7.fit$cluster)
bible.group_testaments<-data.frame(dtm_bnew,km_2.fit$cluster)
```

Analyzing Word Frequencies

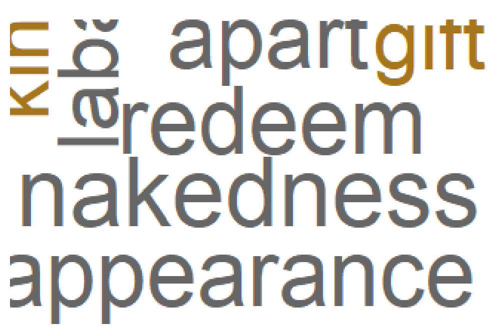
Analysis of word frequencies based on using library package corpus with removing stopwords, stemdocument,numbers, punctuations and finding for the BOOKS

```
#ANALYSIS OF WORD FREQUENCIES FOR 7 SECTIONS
corpus1<-Corpus(VectorSource(bible_st$text));text_corpus1<-tm_map(corpus1,removeWords,my_stopwords1)
## transformation drops documents
text_corpus1 <- tm_map(corpus1,removeWords,my_stopwords2)
## transformation drops documents
text_corpus1 <- tm_map(corpus1, stripWhitespace)
## drops documents
text_corpus1 <- tm_map(corpus1, content_transformer(tolower))
## transformation drops documents
text_corpus1 <- tm_map(corpus1, removeWords, stopwords("english"))
## transformation drops documents
text_corpus1 <- tm_map(corpus1, stemDocument)
## documents
text_corpus1 <- tm_map(corpus1, removeNumbers)
## drops documents
text_corpus1 <- tm_map(corpus1, removePunctuation)
## drops documents
dtm_b2<-DocumentTermMatrix(text_corpus1); dim(dtm_b2)
## [1] 30722 12765
dtm_b221<-removeSparseTerms(dtm_b2,sparse=0.95); dim(dtm_b221)
## [1] 30722    48
dtmr1<-DocumentTermMatrix(text_corpus1,control=list(wordLengths=c(2,20),bounds=list(global=c(2,45))));
## [1] 30722  7454
freq<-sort(colSums(as.matrix(dtmr1)),decreasing = TRUE)
wf1<-data.frame(word=names(freq),freq=freq) ; head(wf1,10)
##                  word freq
## nakedness   nakedness   58
## redeem         redeem   56
## appearance appearance   56
## eateth         eateth   55
## apart           apart   54
## tables         tables   54
## vessel         vessel   52
## salute         salute   52
## sockets       sockets   52
## esther         esther   52
#p1<-ggplot(subset(wf,freq>40),aes(x=reorder(word,freq1),y=freq1))+geom_bar(stat="identity")+
        #    theme(axis.text.x=element_text(angle=45,hjust=1)) #p1
set.seed(142)
```

```
wordcloud(names(freq),freq,min.freq=40,max.words = 100,random.order = FALSE,rot.per = .1,
          random.color=TRUE)
```

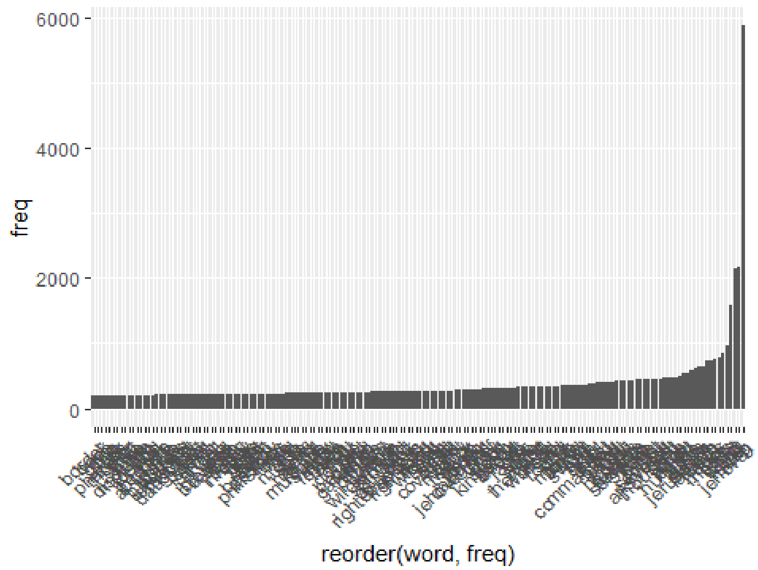



```
wordcloud(names(freq),freq,min.freq=40,max.words = 100,random.order = FALSE,rot.per = .35,
          colors=brewer.pal(8,"Dark2"))
```



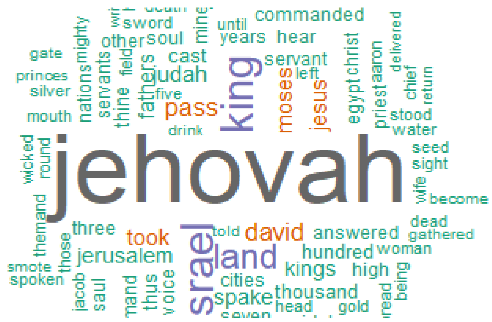

```
#ANALYSIS OF WORD FREQUENCIES FOR 66 BOOKS
corpus<-Corpus(VectorSource(bible_books$text))
text_corpus <- tm_map(corpus,removeWords,my_stopwords1)
## transformation drops documents
text_corpus <- tm_map(corpus,removeWords,my_stopwords2)
## transformation drops documents
text_corpus <- tm_map(corpus, stripWhitespace)
## drops documents
text_corpus <- tm_map(corpus, content_transformer(tolower))
## transformation drops documents
text_corpus <- tm_map(corpus, removeWords, stopwords("english"))
## transformation drops documents
text_corpus <- tm_map(corpus, stemDocument)
## documents
text_corpus <- tm_map(corpus, removeNumbers)
## documents
text_corpus <- tm_map(corpus, removePunctuation)
## drops documents
dtm_b2<-DocumentTermMatrix(text_corpus) ;dim(dtm_b2)
## [1]    66 27727
dtm_b22<-removeSparseTerms(dtm_b2,sparse=0.95) ; dim(dtm_b22);
## [1]   66 5269
```

```r
dtmr <-DocumentTermMatrix(text_corpus, control=list(wordLengths=c(4, 20), bounds = list(global = c
(5,45))))
dim(dtmr) ;
## [1]   66 3965
freq<-sort(colSums(as.matrix(dtmr)),decreasing = TRUE)
wf<-data.frame(word=names(freq),freq=freq); head(wf) ; head(wf,10)
##             word freq
## jehovah jehovah 5870
## king       king 2166
## israel   israel 2150
## land       land 1579
## david     david  972
## pass       pass  843
## moses     moses  769
## took       took  751
## jesus     jesus  737
## judah     judah  723
p<-ggplot(subset(wf,freq>200),aes(x=reorder(word,freq),y=freq))+geom_bar(stat="identity")+
            theme(axis.text.x=element_text(angle=45,hjust=1))
p ; set.seed(142)
```



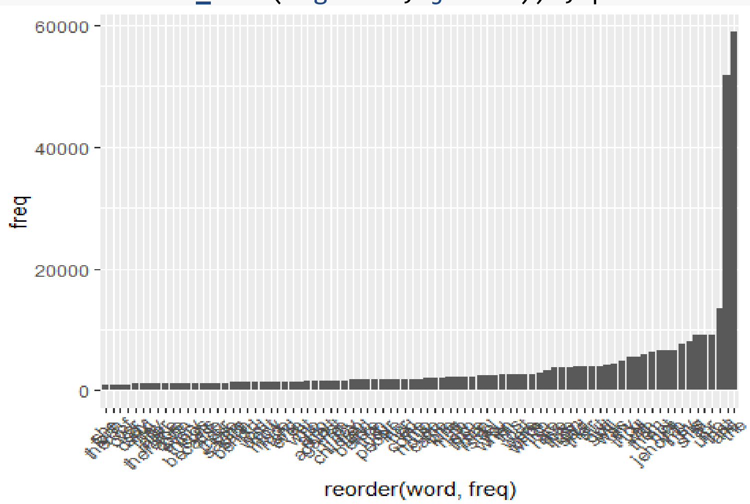

```r
wordcloud(names(freq),freq,min.freq=200,max.words = 100,random.order = FALSE,rot.per = .1,
          random.color=TRUE)
```

```
wordcloud(names(freq),freq,min.freq=200,max.words = 100,random.order = FALSE,rot.per = .35,
          colors=brewer.pal(8,"Dark2"))
```

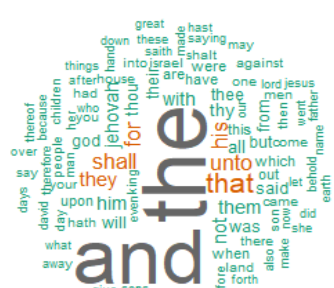



```
#ANALYSIS OF WORD FREQUENCIES FOR 2 TESTAMENTS
corpus<-Corpus(VectorSource(bible_tt$text))
text_corpus <- tm_map(corpus,removeWords,my_stopwords1)
## transformation drops documents
text_corpus <- tm_map(corpus,removeWords,my_stopwords2)
## transformation drops documents
text_corpus <- tm_map(corpus, stripWhitespace)
## drops documents
text_corpus <- tm_map(corpus, content_transformer(tolower))
## transformation drops documents
text_corpus <- tm_map(corpus, removeWords, stopwords("english"))
## transformation drops documents
text_corpus <- tm_map(corpus, stemDocument)
## documents
text_corpus <- tm_map(corpus, removeNumbers)
## documents
text_corpus <- tm_map(corpus, removePunctuation)
## drops documents
dtm_b2<-DocumentTermMatrix(text_corpus);dim(dtm_b2)
## [1] 30722 12765
dtm_b22<-removeSparseTerms(dtm_b2,sparse=0.95);dim(dtm_b22)
## [1] 30722    48
dtmr <-DocumentTermMatrix(text_corpus, control=list(wordLengths=c(2, 20), bounds = list(global = c
(2,45))));dim(dtmr)
## [1] 30722  7454
freq<-sort(colSums(as.matrix(dtmr)),decreasing = TRUE)
wf<-data.frame(word=names(freq),freq=freq); head(wf); head(wf,10)
##                          word freq
## nakedness           nakedness   58
## redeem                 redeem   56
## appearance         appearance   56
## eateth                 eateth   55
## apart                   apart   54
## tables                 tables   54
## vessel                 vessel   52
## salute                 salute   52
## sockets               sockets   52
## esther                 esther   52
## pillar                 pillar   52
## talents               talents   51
```

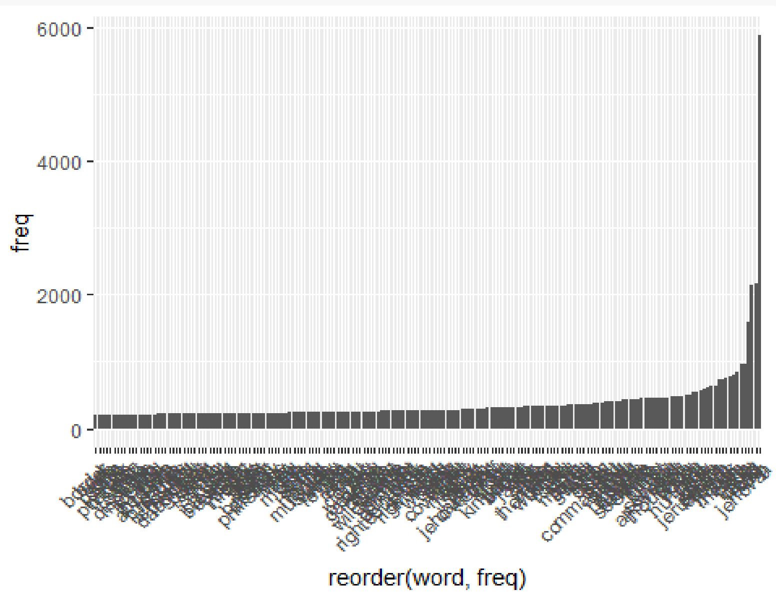
```
#wordcloud(names(freq),freq,min.freq=5,max.words = 10,random.order = FALSE,rot.per = .1,
         # random.color=TRUE)
#wordcloud(names(freq),freq,min.freq=44,max.words = 100,random.order = FALSE,rot.per = .35,
         # colors=brewer.pal(8,"Dark2"))
#ANALYSIS ON WORD FREQUENCY ON THE WHOLE DATASET
freq<-sort(colSums(as.matrix(dtm_b2)),decreasing = TRUE); head(freq,15)
wf<-data.frame(word=names(freq),freq=freq); head(wf,10)
##                  word  freq
## the               the 58738
## and               and 51682
## that             that 13502
## unto             unto  9096
## for               for  9085
## shall           shall  9071
## his               his  8084
## they             they  7569
## jehovah       jehovah  6612
## him               him  6586
## not               not  6543
## them             them  6370
## with             with  5960
p<-ggplot(subset(wf,freq>1000),aes(x=reorder(word,freq),y=freq))+geom_bar(stat="identity")+
            theme(axis.text.x=element_text(angle=45,hjust=1)) ; p
```




```
set.seed(142)
#wordcloud(names(freq),freq,min.freq=866,max.words = 100,random.order = FALSE,rot.per = .1,
 #          random.color=TRUE)
wordcloud(names(freq),freq,min.freq=866,max.words = 100,random.order = FALSE,rot.per = .35,
         colors=brewer.pal(8,"Dark2"))
```



```
#Analysis of the bible_book dataset
corpus<-Corpus(VectorSource(bible_books$text))
text_corpus <- tm_map(corpus,removeWords,my_stopwords1)
## transformation drops documents
text_corpus <- tm_map(corpus,removeWords,my_stopwords2)
## transformation drops documents
text_corpus <- tm_map(corpus, stripWhitespace)
##transformation drops documents
text_corpus <- tm_map(corpus, content_transformer(tolower))
## transformation drops documents
text_corpus <- tm_map(corpus, removeWords, stopwords("english"))
## transformation drops documents
text_corpus <- tm_map(corpus, stemDocument)
## transformation drops documents
text_corpus <- tm_map(corpus, removeNumbers)
## transformation drops documents
text_corpus <- tm_map(corpus, removePunctuation)
## transformation drops documents
dtm_b2<-DocumentTermMatrix(text_corpus);dim(dtm_b2)
## [1]    66 27727
dtm_b22<-removeSparseTerms(dtm_b2,sparse=0.95);dim(dtm_b22)
## [1]   66 5269
dtmr <-DocumentTermMatrix(text_corpus, control=list(wordLengths=c(2, 20), bounds = list(global = c
(2,45))));dim(dtmr)
## [1]    66 10230
freq<-sort(colSums(as.matrix(dtmr)),decreasing = TRUE); head(freq,15)
wf<-data.frame(word=names(freq),freq=freq); head(wf);
##            word freq
## jehovah jehovah 5870
## king       king 2166
## israel   israel 2150
## land       land 1579
## david     david  972
## she         she  966
## pass       pass  843
## two         two  805
## moses     moses  769
## took       took  751
p<-ggplot(subset(wf,freq>200),aes(x=reorder(word,freq),y=freq))+geom_bar(stat="identity")+
            theme(axis.text.x=element_text(angle=45,hjust=1))
p ; set.seed(142)
```



```r
wordcloud(names(freq),freq,min.freq=200,max.words = 100,random.order = FALSE,rot.per = .1,
          random.color=TRUE)
```


```r
wordcloud(names(freq),freq,min.freq=200,max.words = 100,random.order = FALSE,rot.per = .35,
          colors=brewer.pal(8,"Dark2"))
```

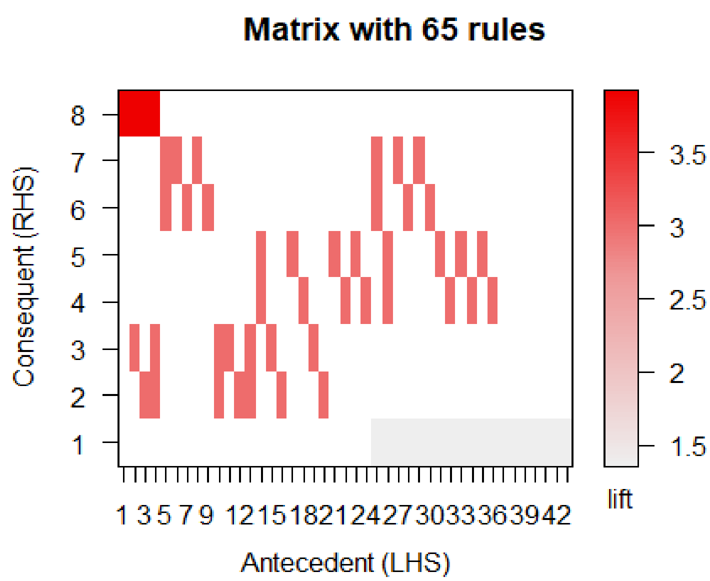
ASSOSICATION RULES

The association rule works good for the books with the rules

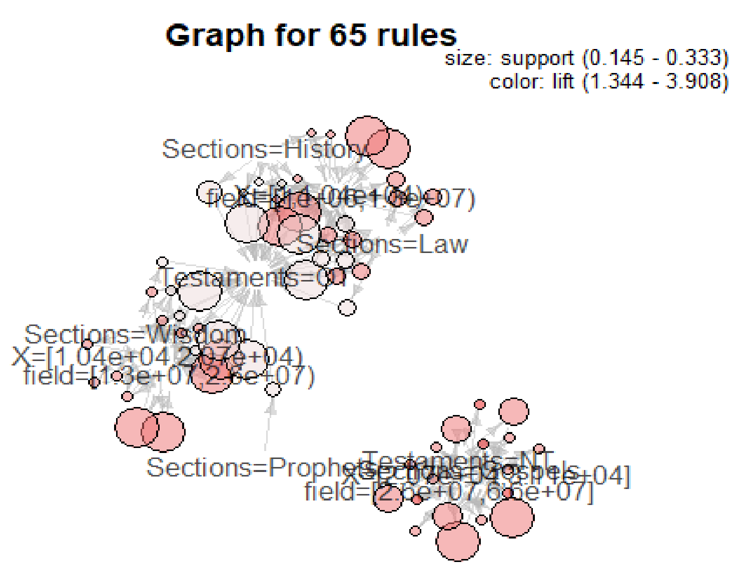
```r
bible_dis<-discretizeDF(bible)
rules_bible<-apriori(bible_dis)
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 3110
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[63095 item(s), 31103 transaction(s)] done [0.08s].
```

```
## sorting and recoding items ... [13 item(s)] done [0.01s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [65 rule(s)] done [0.00s].
## creating S4 object  ... done [0.01s].
summary(rules_bible)
## set of 65 rules
##
## rule length distribution (lhs + rhs):sizes
##  2  3  4
## 23 30 12
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   3.000   2.831   3.000   4.000
##
## summary of quality measures:
##     support         confidence      lift           count
##  Min.   :0.1452   Min.   :1    Min.   :1.344   Min.   : 4516
##  1st Qu.:0.1538   1st Qu.:1    1st Qu.:1.344   1st Qu.: 4785
##  Median :0.1881   Median :1    Median :3.000   Median : 5852
##  Mean   :0.2112   Mean   :1    Mean   :2.572   Mean   : 6568
##  3rd Qu.:0.2559   3rd Qu.:1    3rd Qu.:3.000   3rd Qu.: 7958
##  Max.   :0.3333   Max.   :1    Max.   :3.908   Max.   :10368
##
## mining info:
##       data ntransactions support confidence
##  bible_dis         31103     0.1        0.8
subrules_bible<-rules_bible[quality(rules_bible)$confidence>0.5]
subrules_bible
## set of 65 rules
plot(subrules_bible,method="matrix",measure = "lift")
## Itemsets in Antecedent (LHS)
##  [1] "{X=[2.07e+04,3.11e+04],field=[2.6e+07,6.6e+07],Sections=Gospels}"
##  [2] "{X=[2.07e+04,3.11e+04],Sections=Gospels}"
##  [3] "{field=[2.6e+07,6.6e+07],Sections=Gospels}"
##  [4] "{Sections=Gospels}"
##  [5] "{Testaments=OT,Sections=Wisdom}"
##  [6] "{X=[1.04e+04,2.07e+04),Testaments=OT}"
##  [7] "{field=[1.3e+07,2.6e+07),Testaments=OT}"
##  [8] "{X=[1.04e+04,2.07e+04),Testaments=OT,Sections=Wisdom}"
##  [9] "{field=[1.3e+07,2.6e+07),Testaments=OT,Sections=Wisdom}"
## [10] "{Testaments=NT}"
## [11] "{X=[2.07e+04,3.11e+04]}"
## [12] "{field=[2.6e+07,6.6e+07]}"
## [13] "{Testaments=NT,Sections=Gospels}"
## [14] "{Testaments=OT,Sections=Law}"
## [15] "{X=[2.07e+04,3.11e+04],Testaments=NT}"
## [16] "{field=[2.6e+07,6.6e+07],Testaments=NT}"
## [17] "{X=[1,1.04e+04),Testaments=OT}"
## [18] "{field=[1e+06,1.3e+07),Testaments=OT}"
## [19] "{X=[2.07e+04,3.11e+04],Testaments=NT,Sections=Gospels}"
## [20] "{field=[2.6e+07,6.6e+07],Testaments=NT,Sections=Gospels}"
## [21] "{X=[1,1.04e+04),Testaments=OT,Sections=Law}"
## [22] "{field=[1e+06,1.3e+07),Testaments=OT,Sections=Law}"
## [23] "{X=[1,1.04e+04),Testaments=OT,Sections=History}"
## [24] "{field=[1e+06,1.3e+07),Testaments=OT,Sections=History}"
## [25] "{Sections=Wisdom}"
## [26] "{Sections=Law}"
## [27] "{X=[1.04e+04,2.07e+04)}"
## [28] "{field=[1.3e+07,2.6e+07)}"
## [29] "{X=[1.04e+04,2.07e+04),Sections=Wisdom}"
```

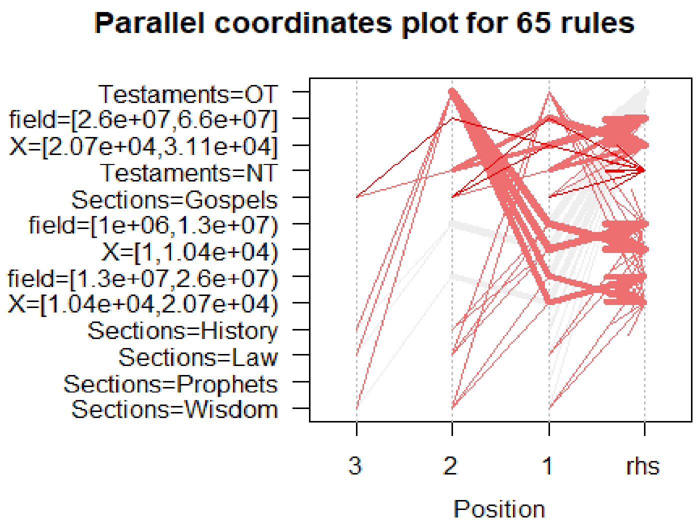
```
## [30] "{field=[1.3e+07,2.6e+07),Sections=Wisdom}"
## [31] "{X=[1,1.04e+04)}"
## [32] "{field=[1e+06,1.3e+07)}"
## [33] "{X=[1,1.04e+04),Sections=Law}"
## [34] "{field=[1e+06,1.3e+07),Sections=Law}"
## [35] "{X=[1,1.04e+04),Sections=History}"
## [36] "{field=[1e+06,1.3e+07),Sections=History}"
## [37] "{Sections=Prophets}"
## [38] "{Sections=History}"
## [39] "{X=[1.04e+04,2.07e+04),field=[1.3e+07,2.6e+07)}"
## [40] "{X=[1,1.04e+04),field=[1e+06,1.3e+07)}"
## [41] "{X=[1.04e+04,2.07e+04),field=[1.3e+07,2.6e+07),Sections=Wisdom}"
## [42] "{X=[1,1.04e+04),field=[1e+06,1.3e+07),Sections=Law}"
## [43] "{X=[1,1.04e+04),field=[1e+06,1.3e+07),Sections=History}"
## Itemsets in Consequent (RHS)
## [1] "{Testaments=OT}"            "{X=[2.07e+04,3.11e+04]}"
## [3] "{field=[2.6e+07,6.6e+07]}" "{X=[1,1.04e+04)}"
## [5] "{field=[1e+06,1.3e+07)}"   "{X=[1.04e+04,2.07e+04)}"
## [7] "{field=[1.3e+07,2.6e+07)}" "{Testaments=NT}"
```



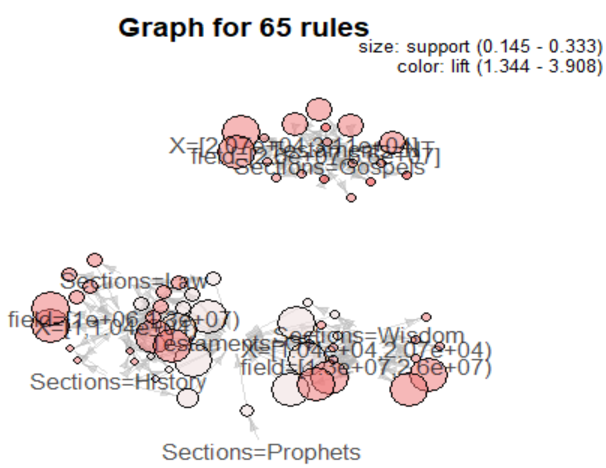

```
subrules_bible2<-head(sort(rules_bible,by="lift"),66)
plot(subrules_bible2,method = "graph")
```



```r
plot(subrules_bible2, method="paracoord")
```



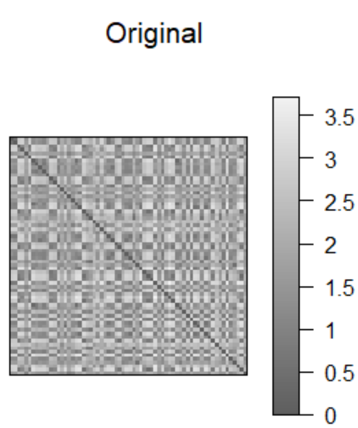

```r
#sel <- plot(rules_bible, measure=c("support", "lift"), shading="confidence", interactive=TRUE)
plot(rules_bible, method="graph")
```



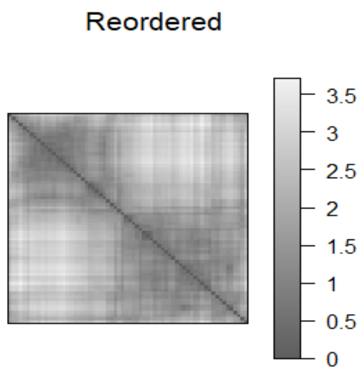
SERATION ANALYSIS

This is the seration analysis for 66 books and ordering according to the seriation analysis based on the document term matrix DTM dataset removing the stopwords and also based on the document-to-document cosine similarity scalar dataset.

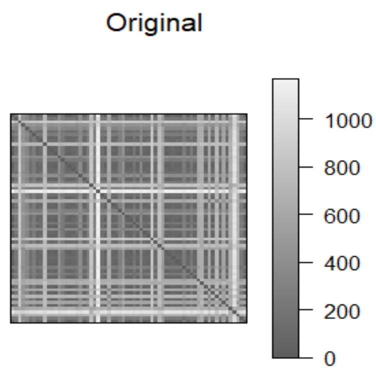
```r
x<-as.matrix(csim_b); x<-x[sample(seq_len(nrow(x))),]
d<-dist(x); o<-seriate(d,method="OLO"); pimage(d,main="Original")
```



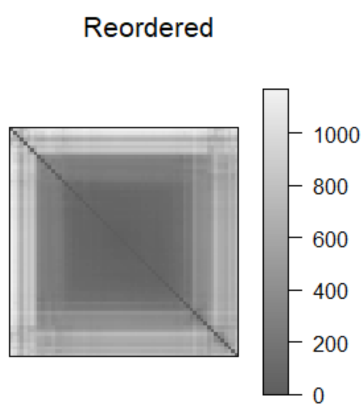

```r
pimage(d,o,main="Reordered")
```




```
get_order(o)
##  [1] 24 64 26 45  4 13 40  3 51 57 31 21 46 39 65 61  6 12 38 44 30 22 59
## [24] 35 23 54 49 42 17 66  5 32 63 36 48  8 10 47  7 19 60 43 33 14 29  9
## [47] 50 52 20 55 37 34 28 41  1  2 56 62 18 25 53 16 27 11 15 58
x1<-as.matrix(dtm_b); x1<-x1[sample(seq_len(nrow(x1))),]
d1<-dist(x1) ; o1<-seriate(d1,method="OLO") ; pimage(d1,main="Original")
```




```
pimage(d1,o1,main="Reordered")
```




```
get_order(o1)
##  [1] 25 63 23 62  3 42 40 54  5 24 31 22  7 49 65  2 19 41 16 59 39 50 47
## [24]  8 30 12 61 48 11 36 27 44 13  1 34 55 57 17 66 52 45 33 21 29 38 18
## [47] 46  9  6 26 51 35 15 32 37 20 14  4 43 58 28 56 60 64 53
```

Based on the seriation analysis the ranking ordered that both the DTM document and the Cosine similarity document have little similar kind of ranking But still when I rerun the output changes each time I run the seriation analysis.