## R Markdown

#**Question 1** #Dataset1 #

a) What is the dimensions and the meaning of the rows and columns of the object dtm\_b? #The dimensions of dtm\_b are 66 1404220,where each row represents the books present in bible and each column represents words that contains in those books

library(textmineR)

library(tidyverse)

library(arules)

library(arulesViz)

library(dplyr)  
library(tm)

library(DT)  
library(ngram)  
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$Testaments[1],1,2))# Creating levels for books as OT =1 & NT =2  
levels(bible$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 levels 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   
  
text.Chapters=c()  
for (i in 1:1189){  
 text.Chapters[i]=paste(text[Chapters==as.character(unique(Chapters)[i])],collapse = "")  
}  
#View(text.Chapters)  
#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  
#Turning the sentences to document term matrix (DTM)  
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)  
#Quest.1 Dimensions of dtm\_b  
dim(dtm\_b)

## [1] 66 1404220

#view(dtm\_b)

#Q2b) Wewouldliketoconsideronlyn-gramsappearingmorethantwotimesinthewholeBible, #andappearing in more than one Book of the Bible. Modify dtm\_b to satisfy this criteria, and call it the same name. For your double checking, the dimension of this modi???ed dtm\_b should be 66-by-24336.

dtm\_b = dtm\_b[,colSums(dtm\_b) >2]  
dtm\_tall = dtm\_b  
dtm\_tall[dtm\_tall >=1]= 1  
dtm\_b = dtm\_tall[,colSums(dtm\_tall) >1]  
dim(dtm\_b)

## [1] 66 23134

#Q1c) Perform Association analyses on the modified dtm\_b in such a way that you can answer: ##Which of these is/are not a frequent itemset with confidence more than 70%: ##{Jerusalem}, #{David}, #{kingdom}, #{covenant}, #{Israel}, #{fear}, #{son}, ##{saith,men,day}, #{spirit,word,man}, #{time,place,make,day,people}, #{heart,things,men,day,children,man}, #{peace,time,evil,word,make,men,day,For,man,God}?

Ans:From the subset rules ,I observed that {saith,men,day},{peace,time,evil,word,make,men,day,For,man,God}, time,place,make,day,people} doesn’t have confidence morethan 70%

ra = as.data.frame(as.matrix(dtm\_b),stringsAsFactors = FALSE)  
tb = discretizeDF(ra, default = list(method = "interval",breaks = 9))  
set = as(tb,"transactions")  
#itemFrequencyPlot(set, topN = 10)  
# Training Apriori on the dataset  
memory.size(max = FALSE)

## [1] 544.97

rules = apriori(data = set, parameter = list(support = 0.94, confidence = 0.7))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.7 0.1 1 none FALSE TRUE 5 0.94 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: 62   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[46268 item(s), 66 transaction(s)] done [0.95s].  
## sorting and recoding items ... [13841 item(s)] done [0.05s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2

## Warning in apriori(data = set, parameter = list(support = 0.94, confidence  
## = 0.7)): Mining stopped (time limit reached). Only patterns up to a length  
## of 2 returned!

## done [6.18s].  
## writing ... [8915527 rule(s)] done [2.36s].  
## creating S4 object ... done [4.97s].

levels(tb$jerusalem)

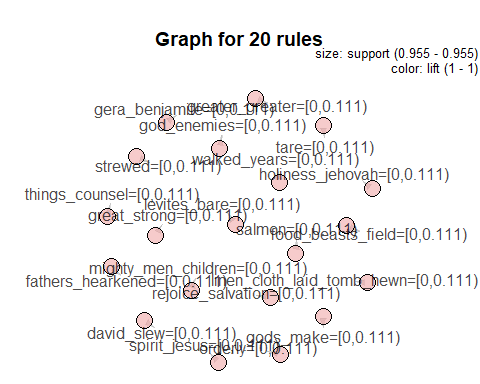
## [1] "[0,0.111)" "[0.111,0.222)" "[0.222,0.333)" "[0.333,0.444)"  
## [5] "[0.444,0.556)" "[0.556,0.667)" "[0.667,0.778)" "[0.778,0.889)"  
## [9] "[0.889,1]"   
arules::inspect((subset(rules,subset=items %in% 'jerusalem=[0.889,1]', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'david=[0.667,0.778)', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'kingdom=[0.889,1]', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'covenant=[0,0.111)', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'israel=[0.889,1]', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'fear=[0.889,1]', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'son=[0,0.111)', by = 'confidence')))  
arules::inspect((subset(rules,subset=items %in% 'son=[0,0.111)', by = 'confidence')))  
  
# Visualising the results  
length(rules)

## [1] 8915527

plot(rules[1:20], method="graph", control=list(type="items"))

## Warning: Unknown control parameters: type

## Available control parameters (with default values):  
## main = Graph for 20 rules



##Q1d) With the response/target variable being Sections, perform Random Forest to find out the top 10 most important variables. Which of these variable(s) is/are not in this list: Christ, faith, land, sea, hope?

ANS:After performing random forest ,from the plot variables such as faith,sea,hope are not in the list

#RANDOMM FOREST  
library(randomForest)

set.seed(123)  
classifier = randomForest(x = ra,  
 y = as.factor(Sections),  
 ntree = 500)  
varImpPlot(classifier,sort = TRUE ,n.var = min(10))

