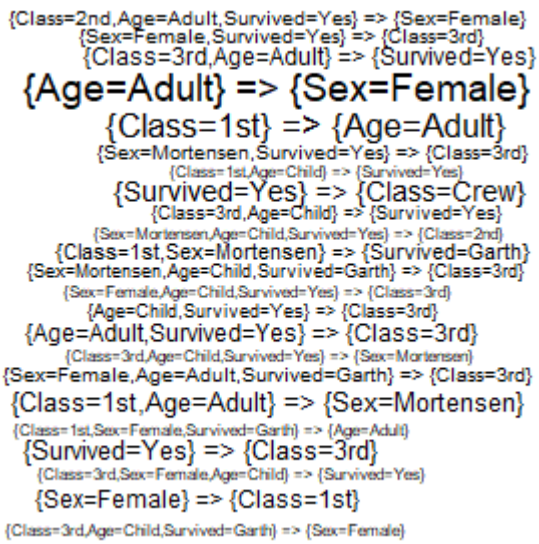
CSIS734-01 Data Mining & Predictive Analytics
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software
CSIS 734-01: Data Mining & Predictive Analytics
Assignment #4 (100 points)
Due Date: March 10th, 2018

Use the "**Titanic_DM.csv**" file on the Canvas for this assignment.

1. Replace "Male" in the csv file with your last name.
   2. Replace "No" in the csv file with your first name.
   3. Set the support threshold to 0.001 and set the confidence threshold to 0.2.
   4. Mine association rules from the new dataset you created in the first two steps.
   5. Print out and capture a screenshot of the **tag cloud** (or text cloud) of the association rules based on the **support**. Your tag cloud should be similar to the following figure (with "Male" being replaced with your last name and "NotSV" being replaced with your first name).
   6. Submit a hardcopy of your answers and screenshots for question 5 in class on the due date. Your hardcopy must be as clear as possible. **Anything that cannot be read won't be graded!**
   7. Please staple all pages of your submission together!!! Instructor is not responsible for missing pages if your submission is not stapled together.
   8. Please also submit your program to clai@stthomas.edu.



```
#CSIS734-01 Data Mining & Predictive Analytics
#Garth Mortensen, mort0052@stthomas.edu


#INSTALL AND LOAD LIBRARIES##############################
#install required packages
#install.packages("arules")
#install.packages("wordcloud")

#Load packages
library(arules)
library(wordcloud)

#LOAD DATA################################################
#create data frame
df <- read.csv("C:/Users/G/Google
Drive/aStThomas/5DataMiningAndPredictiveAnalytics/Assignments/4/Titanic_DM_Replaced.csv",
header = TRUE, stringsAsFactors = TRUE)
#str(df) = #structure(df) #Structure of data
#summary(df)   #Includes 5 most frequent items
#View(df)

#First column to categorical
#categorical variables are called factors in R.
#Because I'm loading categorical variables, I need to first convert them.
#R's default behavior when creating data frames is to convert all characters into factors.
#But column one is integer. See str(df)
#shell.exec("https://stackoverflow.com/questions/49116365/apriori-rules-df-requirements")
df$Passenger <- as.factor(df$Passenger)
#str(df)
#now column 1 is compatible.

#CREATE RULES#############################################
#Step 3 - 4

#create rule in apriori
#Set minimum support (minSup) to 0.001
#Set minimum confidence (minConf) to 0.2
rules <- apriori(df, parameter = list(supp=0.001, conf=0.2))

#Check the results
#inspect(rules)
#inspect(rules[1:20]) #preview the first 20
#View(rules) #this results in a db, not a flat table.
```

```
#WORDCLOUD PREP#########################################
#Base wordcloud on support, so reorder by support.
rules_supp <- sort(rules, by="support", decreasing=TRUE)

#Output ordered rules_supp to csv
write(rules_supp,file = "C:/Users/G/Google
Drive/aStThomas/5DataMiningAndPredictiveAnalytics/Assignments/4/Titanic_DM_Supp.csv",
sep= ",", row.names=FALSE)

#Import them back in to word array
word <- read.csv(file = "C:/Users/G/Google
Drive/aStThomas/5DataMiningAndPredictiveAnalytics/Assignments/4/Titanic_DM_Supp.csv")

#WORDCLOUD################################################
#Step 5

wordcloud(words = word$rules, freq=word$support)


#ETC#####################################################
#These might be useful in the future.

#Basic plot
plot(rules)

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

#Matrix plot of antecedents and consequents
plot(rules[1:20],
    method = "matrix",
    control = list(reorder = TRUE))

#CLEAN UP################################################
rm(list = ls())  #Clear workspace
pacman::p_unload(arules, wordcloud) #Clear packages
dev.off() #Clear plots
cat("\014")  #Clear console (ctrl+L)
```