CSIS734-02 Data Mining & Predictive Analytics
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software
CSIS 734-02: Data Mining & Predictive Analytics
Assignment #2 (100 points)
Due Date: February 24th, 2018

Use the Apriori algorithm to mine association rules from the following transactional database. Let the support threshold be 60% (happened 3 times) and the confidence threshold be 100%.

1. **How many times does the database need to be scanned?**

The scan is used to find candidate itemsets. In this case, the database must be scanned twice.

2. **What are the candidate itemsets for each database scan?**

After scan one, the candidates are {Bread}, {Jelly}, {Milk}, {Peanut Butter} and {Beer}.

After scan two, the candidates are {Bread, Peanut Butter}.

3. **What are the **final** large (frequent) itemsets?**

The final large (frequent) itemset is taken from scan one results, after the support threshold is applied. It includes {Bread}, {Jelly}, {Milk}, {Peanut Butter} and {Beer}. Scan two on candidate 2-itemset {Bread, Peanut Butter} returns a support of only 2/5, or 40%. This is below the support threshold of 3/5, or 60%. Because it does not reach support threshold, no large 2-itemset is created.

4. **What are the final association rules?**

The final association rules are found for large itemsets. The last large itemset to be found was the first, containing {Bread} and {Peanut Butter}. The confidence for its variations are:

Confidence (B -> P) = (B v P) / (B)
= 2 / 4
= 50%

Confidence (P -> B) = (B v P) / (P)
= 2 / 2
= 100%

Key	
B = Bread	
J = Jelly	
M = Milk	
P = Peanut butter	
R = Bread	

TID	Itemset
1	B, J, P
2	B, P
3	B, M, P
4	R, B
5	R, M

Scan ->	Candidate 1-itemset	Support
	{B}	4
	{J}	1
	{M}	2
	{P}	3
	{R}	2

Large 1-itemset	Support
{B}	4
{P}	3

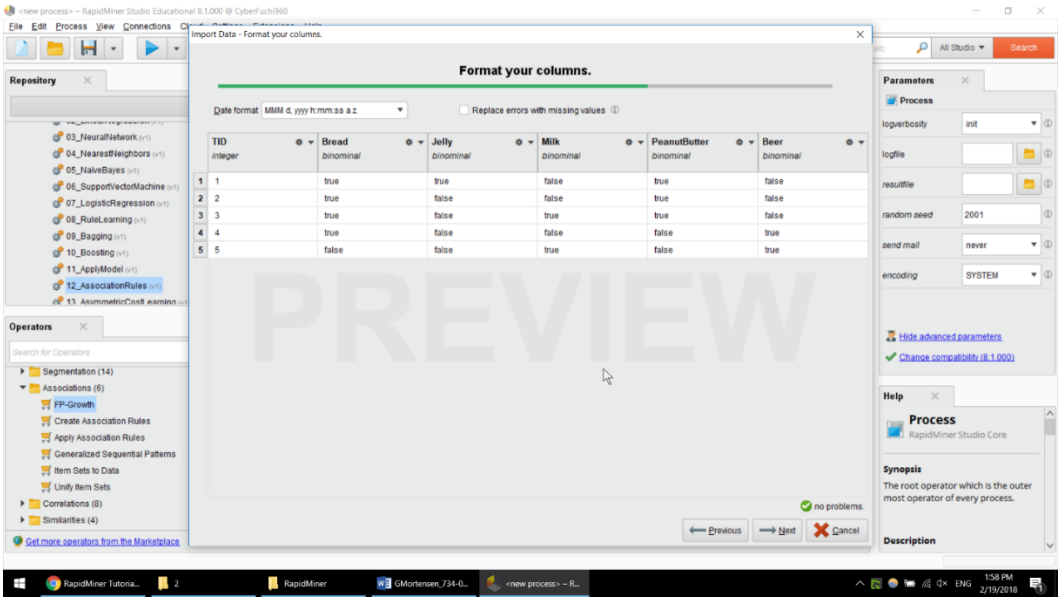
Scan ->	Candidate 2-itemset	Support
	{B, P}	2

Large 2-itemset	Support
---	---

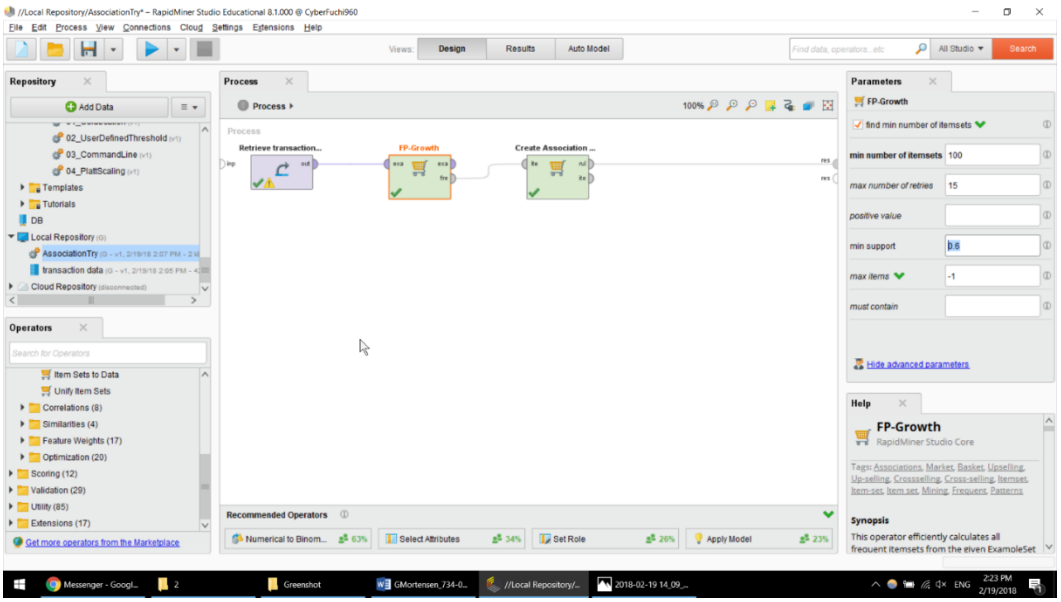
5. <u>(Extra Credits 10 points)</u> Do the following tasks: (1) select your favorite Data Mining tool, (2) convert the following table to the appropriate input format of the tool, (3) run the association rule mining process with the thresholds given in this assignment, (4) get a screen shot of the mining results, (5) submit the screen shot with your answers to question 1 to 4, (6) please also indicate the name of your data mining tool.**

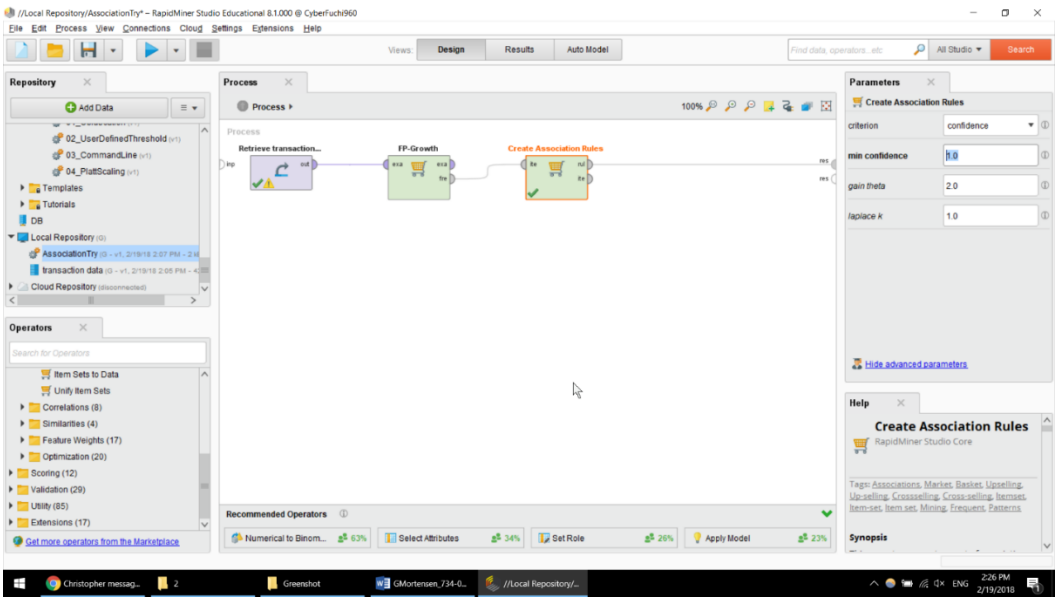
1) RapidMiner

2) convert the following table to the appropriate input format of the tool

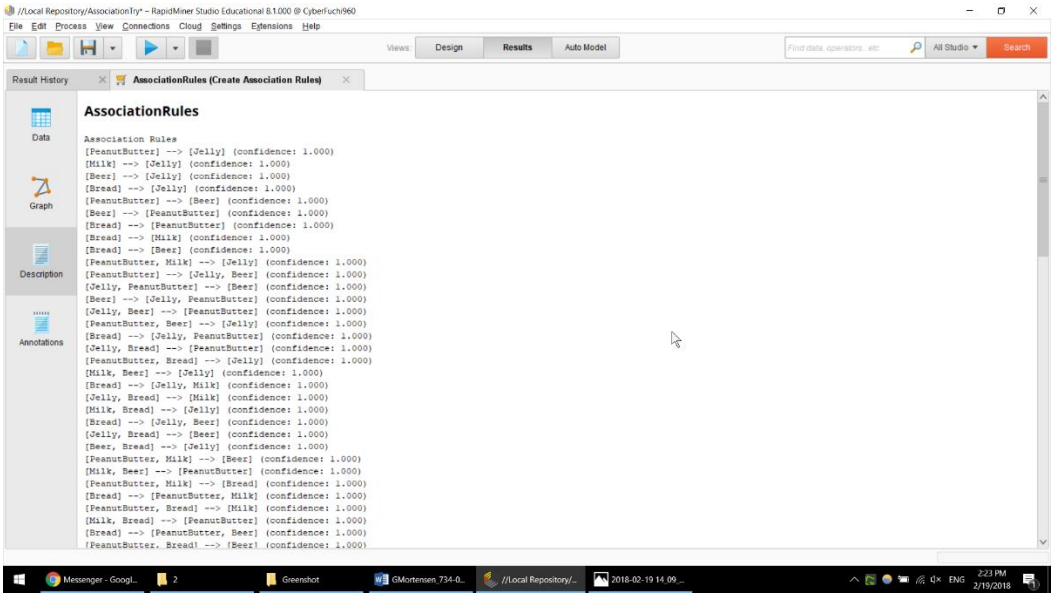


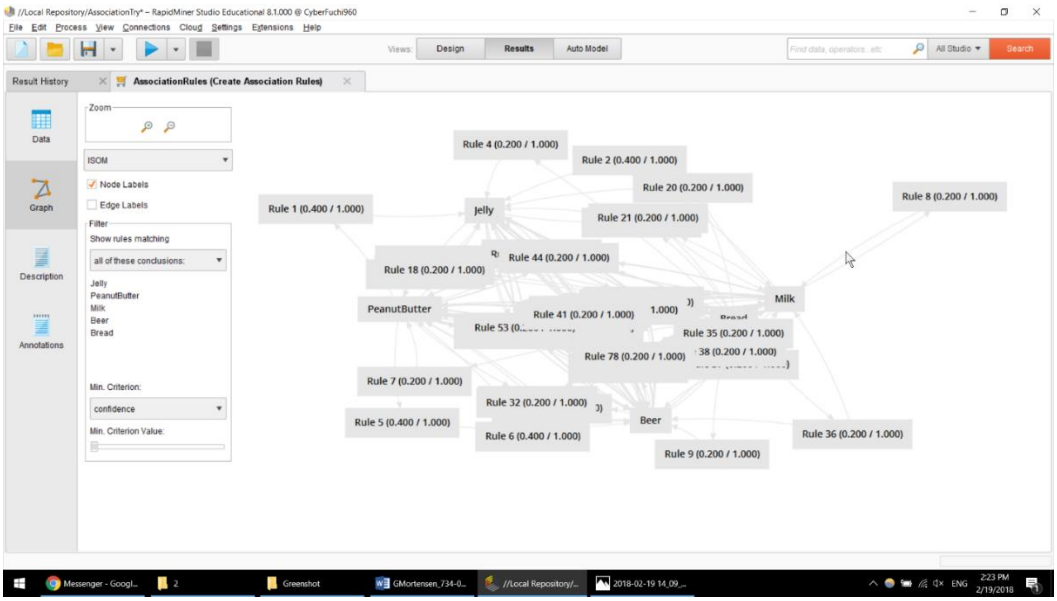
3) run the association rule mining process with the thresholds given in this assignment





4) get a screen shot of the mining results




//Local Repository/AssociationTry* – RapidMiner Studio Educational 8.1.000 @ CyberFuchi960
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Auto Model
Find data, operators, etc.
All Studio
Search
Result History
AssociationRules (Create Association Rules)
Data
Graph
Description
Annotations
Zoom
ISOM
Node Labels
Edge Labels
Filter
Show rules matching:
all of these conclusions:
Jelly
PeanutButter
Milk
Beer
Bread
Min. Criterion:
confidence
Min. Criterion Value:
Rule 4 (0.200 / 1.000)
Rule 2 (0.400 / 1.000)
Rule 20 (0.200 / 1.000)
Rule 8 (0.200 / 1.000)
Rule 1 (0.400 / 1.000)
Jelly
Rule 21 (0.200 / 1.000)
Rule 18 (0.200 / 1.000)
Rule 44 (0.200 / 1.000)
PeanutButter
Rule 41 (0.200 / 1.000)
Milk
Rule 35 (0.200 / 1.000)
Rule 78 (0.200 / 1.000)
Rule 7 (0.200 / 1.000)
Rule 32 (0.200 / 1.000)
Rule 5 (0.400 / 1.000)
Beer
Rule 36 (0.200 / 1.000)
Rule 6 (0.400 / 1.000)
Rule 9 (0.200 / 1.000)
Messenger - Googl...
2
Greenshot
GMortensen_734-0...
//Local Repository/...
2018-02-19 14_09_...
ENG
2:23 PM
2/19/2018

Incomplete script with R:

```
# This is an attempt at association analysis for data mining course.
# it is also my first actual attempt at doing anything more than descriptive statistics with r.

# INSTALL AND LOAD PACKAGES ##################################
pacman::p_load(arules, arulesViz)
# DATA ######################################################
## Read transactional data from arules package
data("Groceries")   # Load data
?Groceries          # Help on data
str(Groceries)      # Structure of data
summary(Groceries)  # Includes 5 most frequent items
# RULES #####################################################
# Set minimum support (minSup) to .001
# Set minimum confidence (minConf) to .75

rules <- apriori(Groceries,
                 parameter = list(supp = 0.001, conf = 0.75))

options(digits=2)
inspect(rules[1:10])

# PLOTS #####################################################
# Scatterplot of support x confidence (colored by lift)
plot(rules)

# Graph of top 20 rules
plot(rules[1:20],
    method = "graph",
```

```
    control = list(type = "items"))

# Parallel coordinates plot of top 20 rules
plot(rules[1:20],
    method = "paracoord",
    control = list(reorder = TRUE))

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

# Grouped matrix plot of antecedents and consequents
plot(rules[1:20], method = "grouped")
```

Note: Assignment will be collected right before the class. **Don't forget to include your name on the top of your answer sheet.**