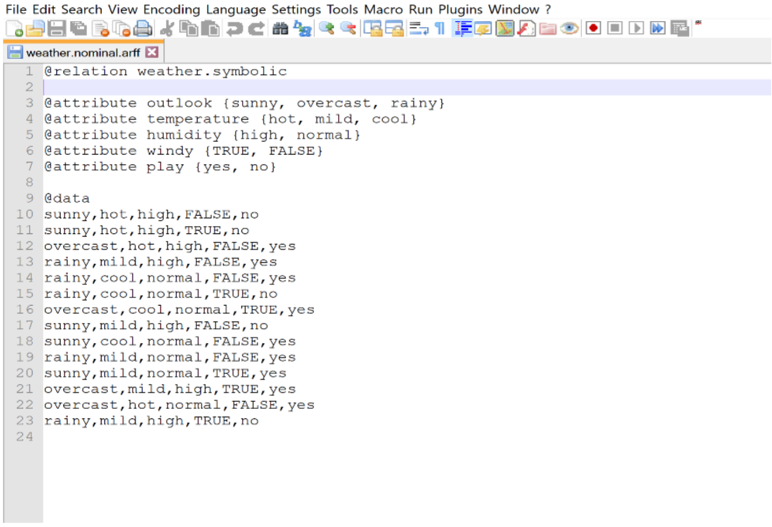
Solution to Laboratory Week 5

1. You will need to have access to the WEKA package.

2. The data files for this lab can be found at /KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data

3. Load the file arff/UCI/weather.nominal.arff.

(a) Inspect the file with an editor.

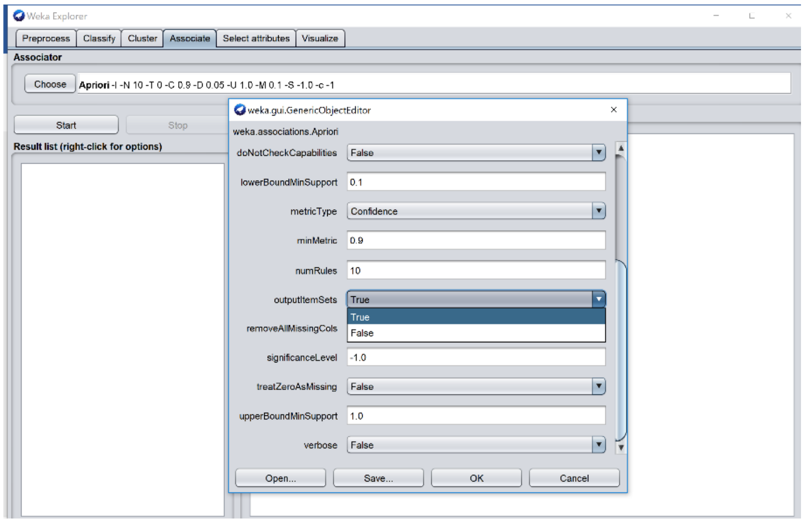

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
weather.nominal.arff

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no

```

Figure 1. Snapshot of view of weather.nominal data set in Notpad++(editor).

As it can be seen from figure1, the data set contains 5 categorical attributes (i.e., 5 columns) and 12 instances (i.e., 12 rows). The last attribute with the name of "play" is the target value and has two classes/labels of yes and no.

(b) Run Apriori using the default settings of the options, but turn on the printing of item sets.



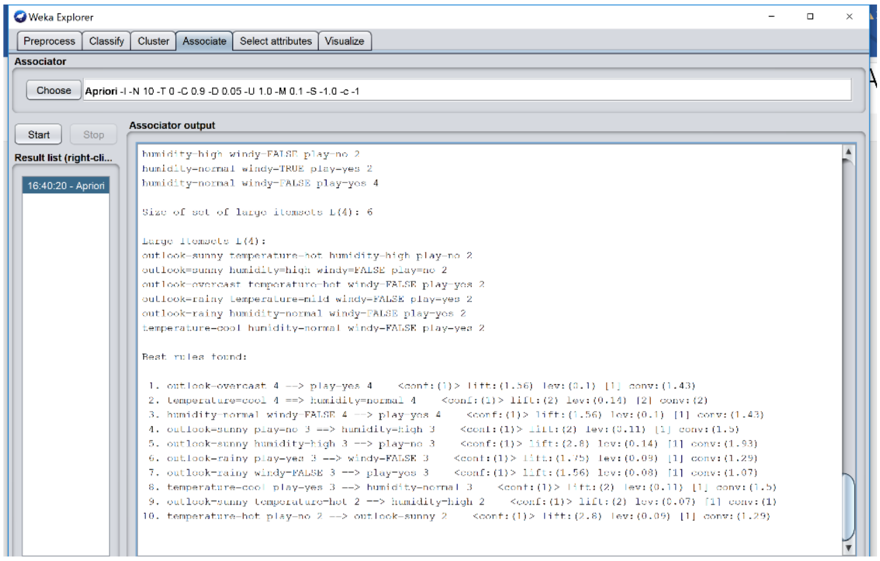
Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

weka.associations.Apriori

doNotCheckCapabilities	False
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.9
numRules	10
outputItemSets	True True False
removeAllMissingCols	
significanceLevel	-1.0
treatZeroAsMissing	False
upperBoundMinSupport	1.0
verbose	False

Figure 2. View of selecting Apriori algorithm from Associate tab and how to turn on the printing of item sets by selecting True on OutputItemSets menu.

(c) What is the confidence for rule 10. How was this confidence value computed?



Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
humidity=high windy=FALSE play=no 2
humidity=normal windy=TRUE play=yes 2
humidity=normal windy=FALSE play=yes 4

Size of set of large itemsets L(4): 6

Large Itemsets L(4):
outlook=sunny temperature=hot humidity=high play=no 2
outlook=sunny humidity=high windy=FALSE play=no 2
outlook=overcast temperature=hot windy=FALSE play=yes 2
outlook=rainy temperature=mild windy=FALSE play=yes 2
outlook=rainy humidity=normal windy=FALSE play=yes 2
temperature=cool humidity=normal windy=FALSE play=yes 2

Best rules found:

 1. outlook=overcast 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
 2. temperature=cool 4 ==> humidity=normal 4    <conf:(1)> lift:(2) lev:(0.14) [2] conv:(2)
 3. humidity=normal windy=FALSE 4 ==> play=yes 4    <conf:(1)> lift:(1.56) lev:(0.1) [1] conv:(1.43)
 4. outlook=sunny play=no 3 ==> humidity=high 3    <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
 5. outlook=sunny humidity=high 3 ==> play=no 3    <conf:(1)> lift:(2.8) lev:(0.14) [1] conv:(1.93)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    <conf:(1)> lift:(1.75) lev:(0.09) [1] conv:(1.29)
 7. outlook=rainy windy=FALSE 3 ==> play=yes 3    <conf:(1)> lift:(1.56) lev:(0.08) [1] conv:(1.07)
 8. temperature=cool play=yes 3 ==> humidity=normal 3    <conf:(1)> lift:(2) lev:(0.11) [1] conv:(1.5)
 9. outlook=sunny temperature=hot 2 ==> humidity=high 2    <conf:(1)> lift:(2) lev:(0.07) [1] conv:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2    <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)
```

Figure 3. Output from Run of Apriori on weather.nominal data set.

Below is rule 10, where its confidence is 1 as shown conf:(1).

Rule : "10. temperature=hot play=no 2 ==> outlook=sunny 2   <conf:(1)> lift:(2.8) lev:(0.09) [1] conv:(1.29)"

Support is 2 or 2/14 or 14.2%

Confidence is 2/2 = 1 or 100%

(d) How many instances form the support of rule 8?

3 instances.

(e) Identify the itemset from which rule 1 was generated. What other rules could be generated from this itemset?

This is rule 1 : outlook=overcast 4 ==> play=yes 4.

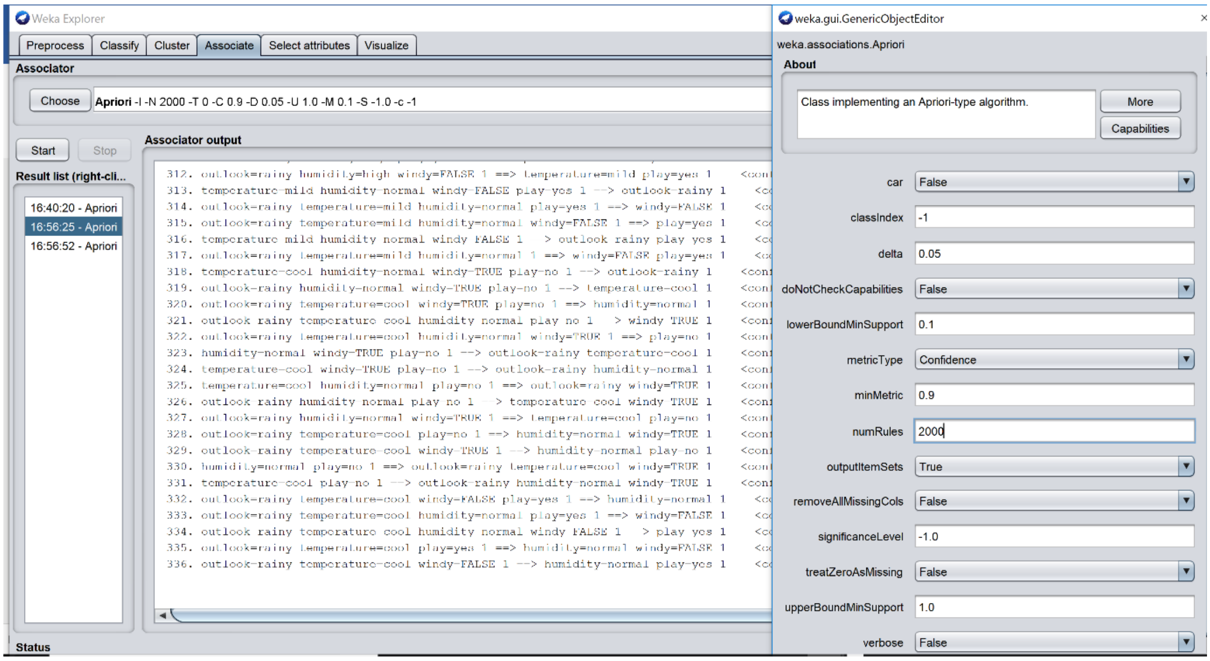
The itemset that the rule 1 has been generated from is a set contains outlook=overcast and play= yes items (i.e., set of rule 1 = {outlook=overcast, play= yes}.

(f) What does 'best rules' mean? What criteria are used to determine the best rules?

The best rules are rules with the highest support and high confidence (i.e., confidence equal or close to 1).

(g) How many rules can be generated from this data? Experiment with the numRules parameter.

As shown in figure 4, by setting numRules to 2000, the maximum number of rules that will pull out by running Apriori on top of the data set is 336 rules. You can try with smaller or larger number than 336 to see the difference. If you set number of rules to any number greater than 336, it would only return 336 rules not any more.



```
Apriori -I -N 2000 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
```

Associator output

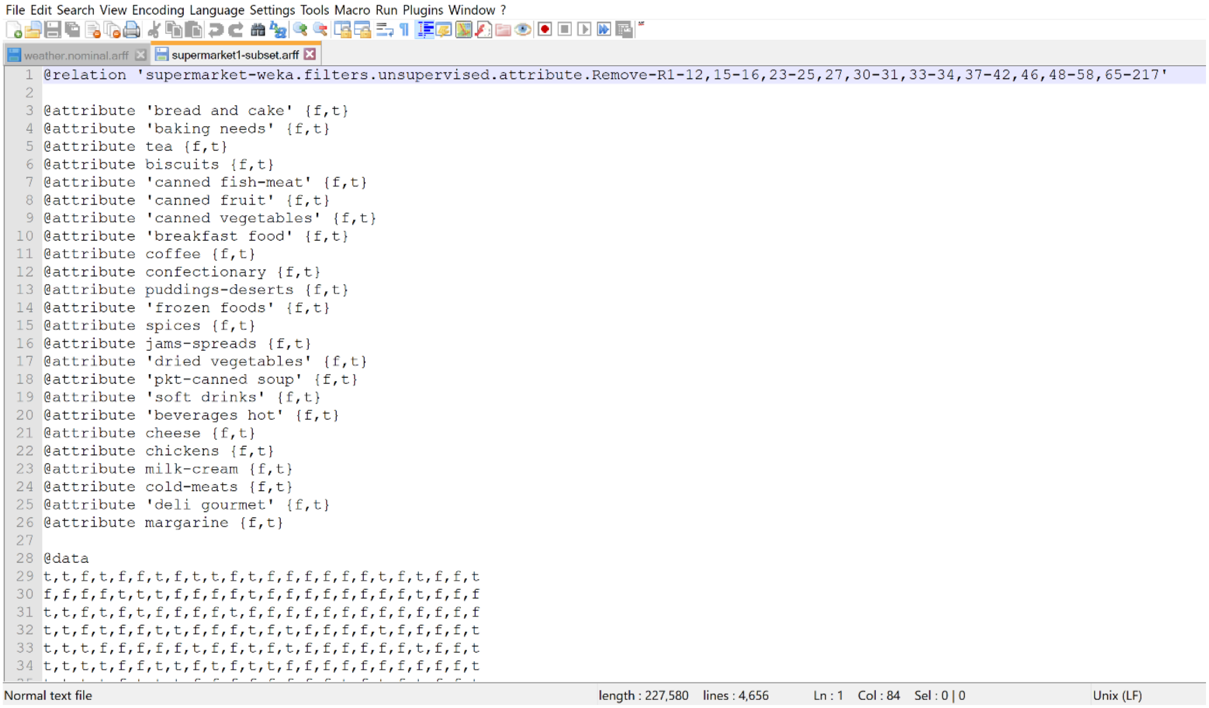
```
312. outlook=rainy humidity=high windy=FALSE 1 ==> temperature=mild play=yes 1    <con
313. temperature=mild humidity=normal windy=FALSE play=yes 1 ==> outlook=rainy 1    <co
314. outlook=rainy temperature=mild humidity=normal play=yes 1 ==> windy=FALSE 1    <co
315. outlook=rainy temperature=mild humidity=normal windy=FALSE 1 ==> play=yes 1    <co
316. temperature=mild humidity=normal windy=FALSE 1 ==> outlook=rainy play=yes 1    <co
317. outlook=rainy temperature=mild humidity=normal 1 ==> windy=FALSE play=yes 1    <co
318. temperature=cool humidity=normal windy=TRUE play=no 1 ==> outlook=rainy 1    <con
319. outlook=rainy humidity=normal windy=TRUE play=no 1 ==> temperature=cool 1    <con
320. outlook=rainy temperature=cool windy=TRUE play=no 1 ==> humidity=normal 1    <con
321. outlook=rainy temperature=cool humidity=normal play=no 1 ==> windy=TRUE 1    <con
322. outlook=rainy temperature=cool humidity=normal windy=TRUE 1 ==> play=no 1    <con
323. humidity=normal windy=TRUE play=no 1 ==> outlook=rainy temperature=cool 1    <con
324. temperature=cool windy=TRUE play=no 1 ==> outlook=rainy humidity=normal 1    <con
325. temperature=cool humidity=normal play=no 1 ==> outlook=rainy windy=TRUE 1    <con
326. outlook=rainy humidity=normal play=no 1 ==> temperature=cool windy=TRUE 1    <con
327. outlook=rainy humidity=normal windy=TRUE 1 ==> temperature=cool play=no 1    <con
328. outlook=rainy temperature=cool play=no 1 ==> humidity=normal windy=TRUE 1    <con
329. outlook=rainy temperature=cool windy=TRUE 1 ==> humidity=normal play=no 1    <con
330. humidity=normal play=no 1 ==> outlook=rainy temperature=cool windy=TRUE 1    <con
331. temperature=cool play=no 1 ==> outlook=rainy humidity=normal windy=TRUE 1    <con
332. outlook=rainy temperature=cool windy=FALSE play=yes 1 ==> humidity=normal 1    <co
333. outlook=rainy temperature=cool humidity=normal play=yes 1 ==> windy=FALSE 1    <co
334. outlook=rainy temperature=cool humidity=normal windy=FALSE 1 ==> play=yes 1    <co
335. outlook=rainy temperature=cool play=yes 1 ==> humidity=normal windy=FALSE 1    <co
336. outlook=rainy temperature=cool windy=FALSE 1 ==> humidity=normal play=yes 1    <co
```

car: False
classIndex: -1
delta: 0.05
doNotCheckCapabilities: False
lowerBoundMinSupport: 0.1
metricType: Confidence
minMetric: 0.9
numRules: 2000
outputItemSets: True
removeAllMissingCols: False
significanceLevel: -1.0
treatZeroAsMissing: False
upperBoundMinSupport: 1.0
verbose: False

Figure 4. Maximum number of rules that pules out from running apriori on weather data set is 336 even by setting parameter numRules= 2000.

4. Load the file arff/supermarket1-subset.arff.

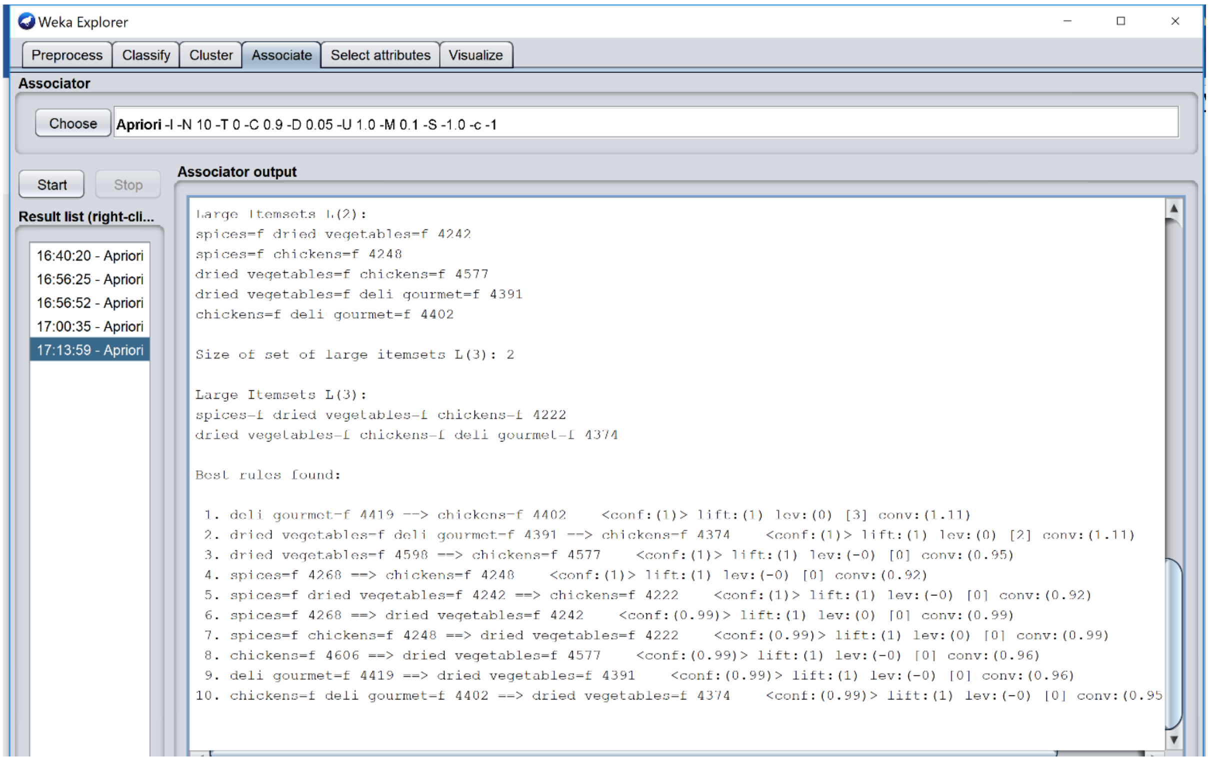
(a) View the file with an editor.


File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
weather.nominal.arff
supermarket1-subset.arff
Normal text file
length : 227,580   lines : 4,656
Ln : 1   Col : 84   Sel : 0 | 0
Unix (LF)

```
@relation 'supermarket-weka.filters.unsupervised.attribute.Remove-R1-12,15-16,23-25,27,30-31,33-34,37-42,46,48-58,65-217'

@attribute 'bread and cake' {f,t}
@attribute 'baking needs' {f,t}
@attribute tea {f,t}
@attribute biscuits {f,t}
@attribute 'canned fish-meat' {f,t}
@attribute 'canned fruit' {f,t}
@attribute 'canned vegetables' {f,t}
@attribute 'breakfast food' {f,t}
@attribute coffee {f,t}
@attribute confectionary {f,t}
@attribute puddings-deserts {f,t}
@attribute 'frozen foods' {f,t}
@attribute spices {f,t}
@attribute jams-spreads {f,t}
@attribute 'dried vegetables' {f,t}
@attribute 'pkt-canned soup' {f,t}
@attribute 'soft drinks' {f,t}
@attribute 'beverages hot' {f,t}
@attribute cheese {f,t}
@attribute chickens {f,t}
@attribute milk-cream {f,t}
@attribute cold-meats {f,t}
@attribute 'deli gourmet' {f,t}
@attribute margarine {f,t}

@data
t,t,f,t,f,f,t,f,t,t,f,t,f,f,f,f,f,f,t,f,t,f,f,t
f,f,f,f,t,t,t,f,f,f,f,t,f,f,f,f,f,f,f,f,t,f,f,f
t,t,f,t,f,t,f,f,f,f,t,f,f,f,f,f,f,f,f,f,f,f,f,f
t,t,f,t,f,f,t,t,f,f,f,t,f,t,f,f,f,f,t,f,f,f,f,t
t,t,t,f,f,f,f,f,t,f,f,t,f,t,f,f,f,f,f,f,t,f,f,t
t,t,t,t,f,f,t,t,f,t,f,t,t,f,f,f,f,f,f,f,f,f,f,t
```

Figure 5. View of the data set in Notepad++ editor.

(b) Go to the Associate screen and run Apriori with the default values.



Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

16:40:20 - Apriori
16:56:25 - Apriori
16:56:52 - Apriori
17:00:35 - Apriori
17:13:59 - Apriori

```
Large Itemsets L(2):
spices=f dried vegetables=f 4242
spices=f chickens=f 4248
dried vegetables=f chickens=f 4577
dried vegetables=f deli gourmet=f 4391
chickens=f deli gourmet=f 4402

Size of set of large itemsets L(3): 2

Large Itemsets L(3):
spices=f dried vegetables=f chickens=f 4222
dried vegetables=f chickens=f deli gourmet=f 4374

Best rules found:

 1. deli gourmet=f 4419 ==> chickens=f 4402    <conf:(1)> lift:(1) lev:(0) [3] conv:(1.11)
 2. dried vegetables=f deli gourmet=f 4391 ==> chickens=f 4374    <conf:(1)> lift:(1) lev:(0) [2] conv:(1.11)
 3. dried vegetables=f 4598 ==> chickens=f 4577    <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.95)
 4. spices=f 4268 ==> chickens=f 4248    <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.92)
 5. spices=f dried vegetables=f 4242 ==> chickens=f 4222    <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.92)
 6. spices=f 4268 ==> dried vegetables=f 4242    <conf:(0.99)> lift:(1) lev:(0) [0] conv:(0.99)
 7. spices=f chickens=f 4248 ==> dried vegetables=f 4222    <conf:(0.99)> lift:(1) lev:(0) [0] conv:(0.99)
 8. chickens=f 4606 ==> dried vegetables=f 4577    <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.96)
 9. deli gourmet=f 4419 ==> dried vegetables=f 4391    <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.96)
10. chickens=f deli gourmet=f 4402 ==> dried vegetables=f 4374    <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.95
```

Figure 6. Snapshot of outputs of running Apriori on supermarket-subset dataset.

(c) Are there any rules with high accuracy?

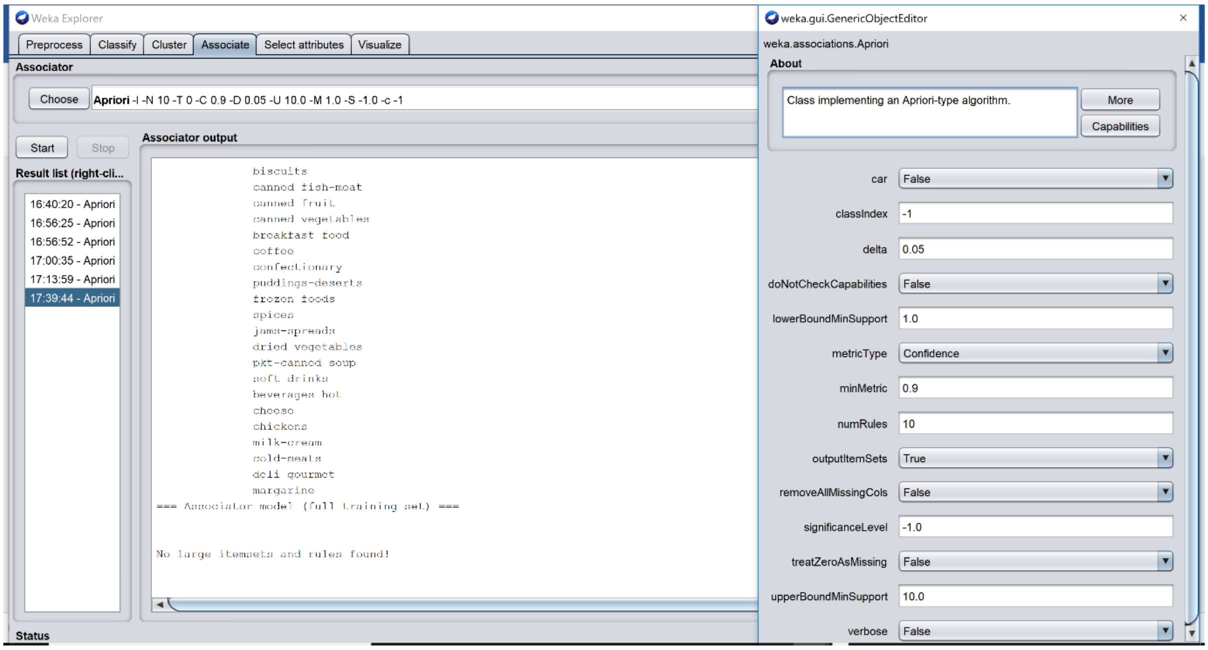
Many

(d) Are there any golden nuggets in the output?

There are many strong rules, however the relationships captured are of the form: not buying some items results in not buying other items. These relationships are not useful in practice and are unlikely to yield golden nuggets.

(Questions e, f, g are editing mistakes.)

(h) Experiment with different metrics. How do the generated rules change with the different metrics.



Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 10.0 -M 1.0 -S -1.0 -c -1

16:40:20 - Apriori
16:56:25 - Apriori
16:56:52 - Apriori
17:00:35 - Apriori
17:13:59 - Apriori
17:39:44 - Apriori

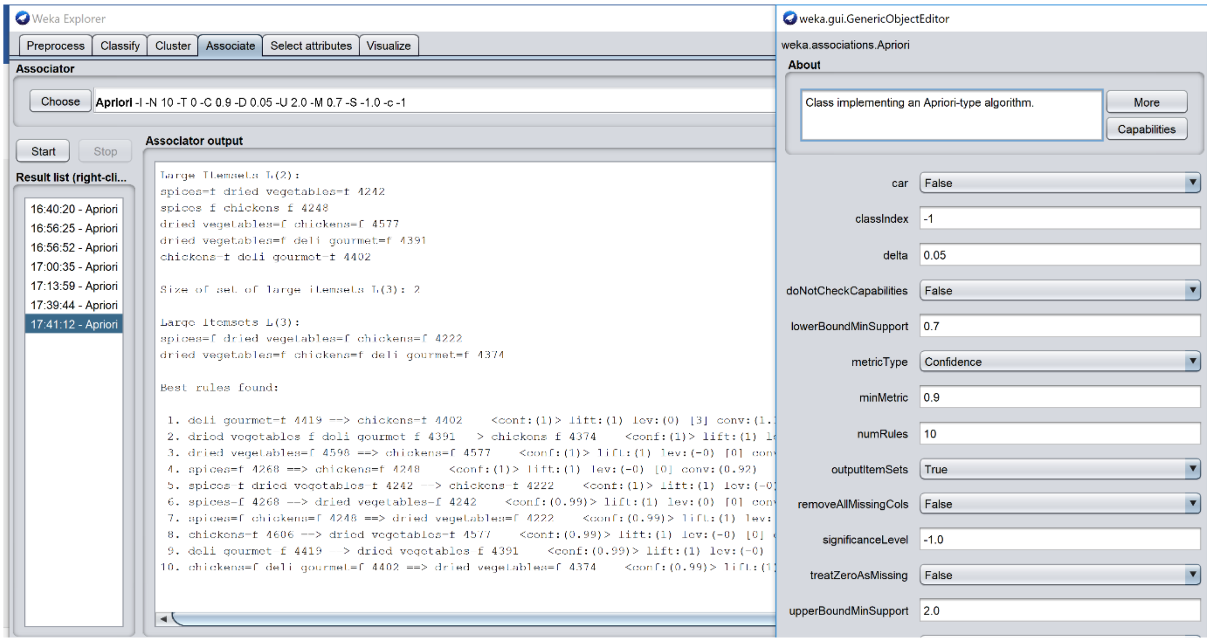
Associator output

```
              biscuits
              canned fish-meat
              canned fruit
              canned vegetables
              breakfast food
              coffee
              confectionary
              puddings-deserts
              frozen foods
              spices
              jams-spreads
              dried vegetables
              pkt-canned soup
              soft drinks
              beverages hot
              cheese
              chickens
              milk-cream
              cold-meats
              deli gourmet
              margarine
=== Associator model (full training set) ===


No large itemsets and rules found!
```

weka.associations.Apriori

Class implementing an Apriori-type algorithm.

car: False
classIndex: -1
delta: 0.05
doNotCheckCapabilities: False
lowerBoundMinSupport: 1.0
metricType: Confidence
minMetric: 0.9
numRules: 10
outputItemSets: True
removeAllMissingCols: False
significanceLevel: -1.0
treatZeroAsMissing: False
upperBoundMinSupport: 10.0
verbose: False



Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 2.0 -M 0.7 -S -1.0 -c -1

16:40:20 - Apriori
16:56:25 - Apriori
16:56:52 - Apriori
17:00:35 - Apriori
17:13:59 - Apriori
17:39:44 - Apriori
17:41:12 - Apriori

Associator output

```
Large Itemsets L(2):
spices=t dried vegetables=t 4242
spices f chickens f 4248
dried vegetables=f chickens=f 4577
dried vegetables=f deli gourmet=f 4391
chickens=t deli gourmet=t 4402

Size of set of large itemsets L(3): 2

Large Itemsets L(3):
spices=f dried vegetables=f chickens=f 4222
dried vegetables=f chickens=f deli gourmet=f 4374

Best rules found:

 1. deli gourmet=t 4419 ==> chickens=t 4402    <conf:(1)> lift:(1) lev:(0) [3] conv:(1.
 2. dried vegetables f deli gourmet f 4391   > chickens f 4374    <conf:(1)> lift:(1) l
 3. dried vegetables=f 4598 ==> chickens=f 4577    <conf:(1)> lift:(1) lev:(-0) [0] con
 4. spices=f 4268 ==> chickens=f 4248    <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.92)
 5. spices t dried vegetables t 4242   > chickens t 4222    <conf:(1)> lift:(1) lev:(-0
 6. spices=f 4268 ==> dried vegetables=f 4242    <conf:(0.99)> lift:(1) lev:(0) [0] con
 7. spices=f chickens=f 4248 ==> dried vegetables=f 4222    <conf:(0.99)> lift:(1) lev:
 8. chickens=t 4606 ==> dried vegetables=t 4577    <conf:(0.99)> lift:(1) lev:(-0) [0] 
 9. deli gourmet f 4419   > dried vegetables f 4391    <conf:(0.99)> lift:(1) lev:(-0)
10. chickens=f deli gourmet=f 4402 ==> dried vegetables=f 4374    <conf:(0.99)> lift:(1
```

weka.associations.Apriori

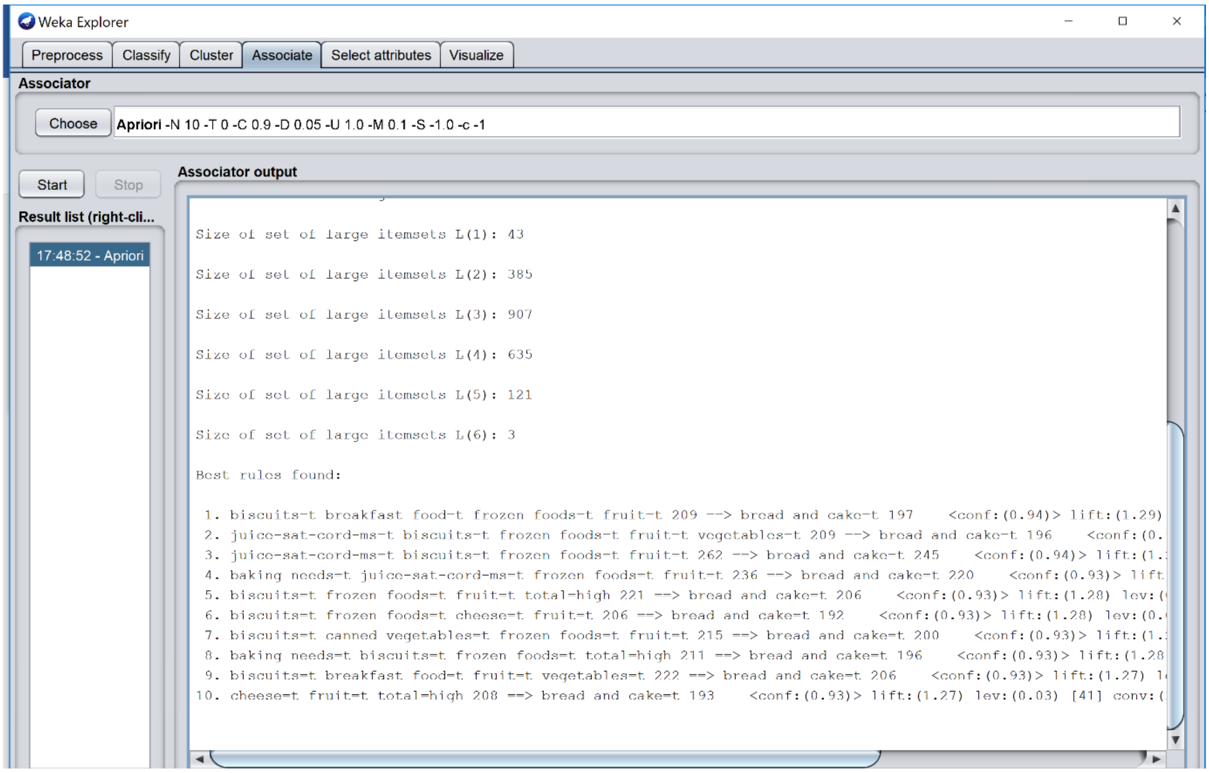
Class implementing an Apriori-type algorithm.

car: False
classIndex: -1
delta: 0.05
doNotCheckCapabilities: False
lowerBoundMinSupport: 0.7
metricType: Confidence
minMetric: 0.9
numRules: 10
outputItemSets: True
removeAllMissingCols: False
significanceLevel: -1.0
treatZeroAsMissing: False
upperBoundMinSupport: 2.0

Figure 7. Effect of changing lowerBoundMinSupport and UpperBoundMinSupport.

5. Load the file arff/supermarket2-small.arff.

(a) Repeat (4) above on this data.



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
Size of set of large itemsets L(1): 43

Size of set of large itemsets L(2): 385

Size of set of large itemsets L(3): 907

Size of set of large itemsets L(4): 635

Size of set of large itemsets L(5): 121

Size of set of large itemsets L(6): 3

Best rules found:

 1. biscuits=t breakfast food=t frozen foods=t fruit=t 209 ==> bread and cake=t 197    <conf:(0.94)> lift:(1.29)
 2. juice-sat-cord-ms=t biscuits=t frozen foods=t fruit=t vegetables=t 209 ==> bread and cake=t 196    <conf:(0.
 3. juice-sat-cord-ms=t biscuits=t frozen foods=t fruit=t 262 ==> bread and cake=t 245    <conf:(0.94)> lift:(1.
 4. baking needs=t juice-sat-cord-ms=t frozen foods=t fruit=t 236 ==> bread and cake=t 220    <conf:(0.93)> lift
 5. biscuits=t frozen foods=t fruit=t total=high 221 ==> bread and cake=t 206    <conf:(0.93)> lift:(1.28) lev:(
 6. biscuits=t frozen foods=t cheese=t fruit=t 206 ==> bread and cake=t 192    <conf:(0.93)> lift:(1.28) lev:(0.
 7. biscuits=t canned vegetables=t frozen foods=t fruit=t 215 ==> bread and cake=t 200    <conf:(0.93)> lift:(1.
 8. baking needs=t biscuits=t frozen foods=t total=high 211 ==> bread and cake=t 196    <conf:(0.93)> lift:(1.28
 9. biscuits=t breakfast food=t fruit=t vegetables=t 222 ==> bread and cake=t 206    <conf:(0.93)> lift:(1.27) le
10. cheese=t fruit=t total=high 208 ==> bread and cake=t 193    <conf:(0.93)> lift:(1.27) lev:(0.03) [41] conv:(
```

Figure 8. Output of running Apriori on supermarket2-small data set.s

(b) What do you conclude about the different representations?

In previous data set all values are represented by true or false as "t" (i.e., purchased) and "f"(i.e., not purchased) and apriori generated rules from not purchased items which were not very useful. In this data set, values are represented by either "t" or "?" (i.e., missing values) and the generated rules are between purchased items. Most of the strongest rules concern bread and cake. I don't know enough about supermarket operations to know whether this is a golden nugget or not.

6. Repeat (4) and (5) with FilteredAssociator and FPGrowth and compare the results with apriori.

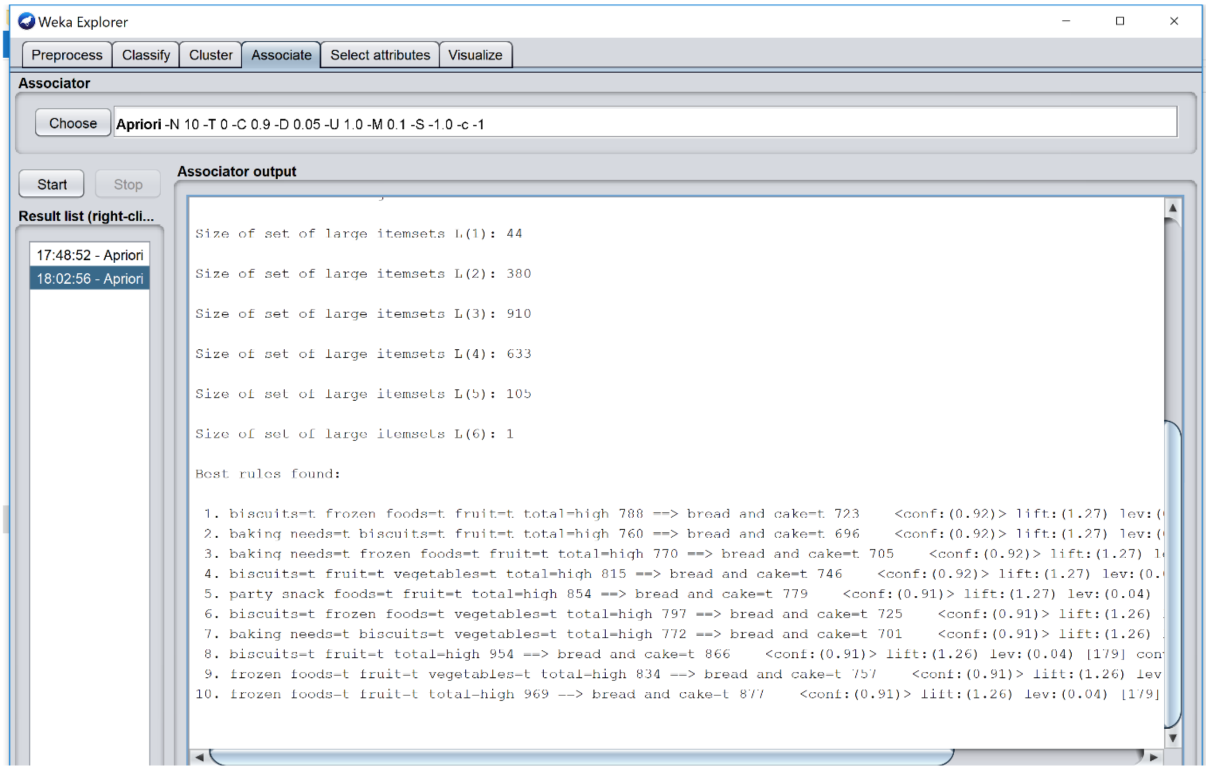
7. The files supermarket1-subset.arff and supermarket2-small.arff are trimmed down versions of supermarket1.arff and supermarket2.arff in order to get quick execution times for apriori.

(a) Run apriori on the larger files.

There are memory and computation time problems. How long should we wait for the run to finish? Hours? Days? Weeks?

(b) What problems arise and what can you do about them?

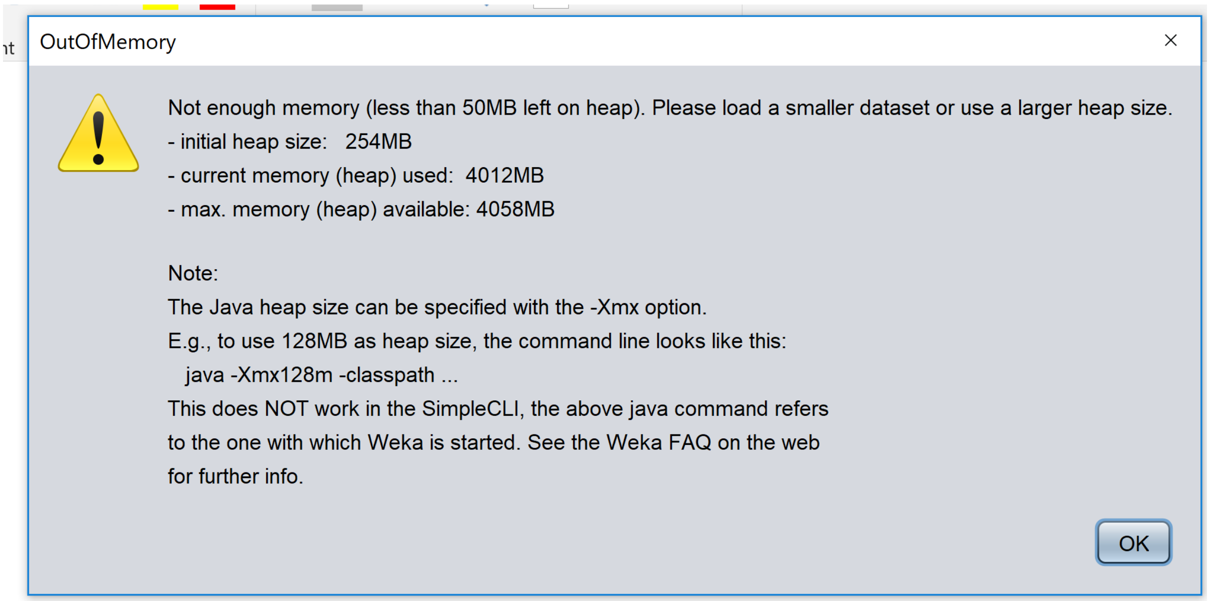
We need to experiment with smaller numbers of instances and/or attributes to get a sense of how the algorithms scale with increasing amounts of data.



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
Size of set of large itemsets L(1): 44

Size of set of large itemsets L(2): 380

Size of set of large itemsets L(3): 910

Size of set of large itemsets L(4): 633

Size of set of large itemsets L(5): 105

Size of set of large itemsets L(6): 1

Best rules found:

 1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723    <conf:(0.92)> lift:(1.27) lev:(
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696    <conf:(0.92)> lift:(1.27) lev:(
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705    <conf:(0.92)> lift:(1.27) l
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746    <conf:(0.92)> lift:(1.27) lev:(0.
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779    <conf:(0.91)> lift:(1.27) lev:(0.04)
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725    <conf:(0.91)> lift:(1.26)
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701    <conf:(0.91)> lift:(1.26)
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] con
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757    <conf:(0.91)> lift:(1.26) lev
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877    <conf:(0.91)> lift:(1.26) lev:(0.04) [179]
```

Figure 9. Running Apriori on supermarket2.



Not enough memory (less than 50MB left on heap). Please load a smaller dataset or use a larger heap size.
- initial heap size: 254MB
- current memory (heap) used: 4012MB
- max. memory (heap) available: 4058MB

Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers
to the one with which Weka is started. See the Weka FAQ on the web
for further info.

Figure 10. Running Apriori on supermarket1.