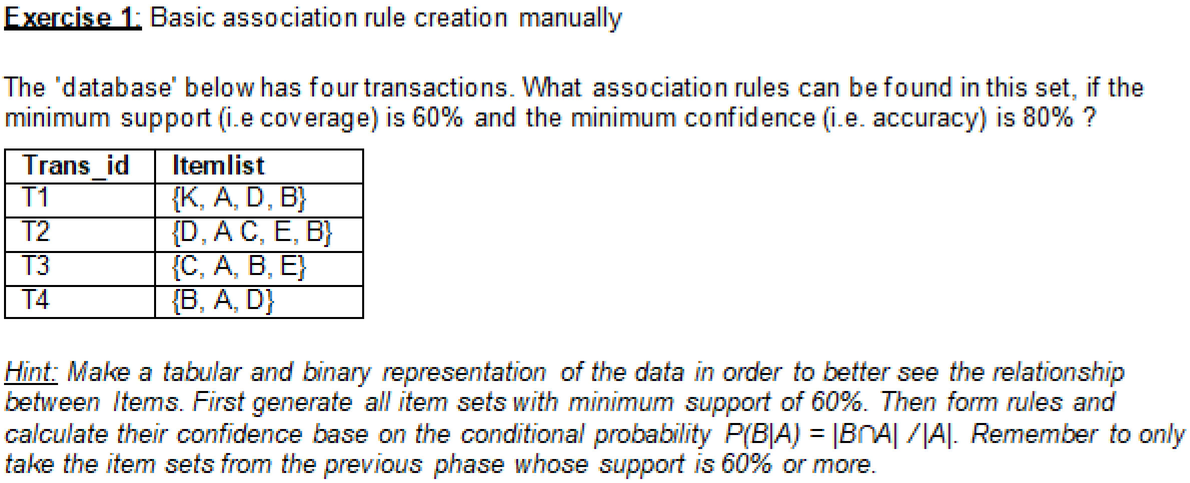
Name: Tejal Yadav
Branch: Computer
Year: Third year
UID: 2019130071
Subject: Data Analytics
Date: 17th May 2022

Experiment 5: Apriori Algorithm and Association rule mining with WEKA

Objective: To apply Apriori Algorithm to given dataset and perform Association Rule Mining with WEKA



Exercise 1: Basic association rule creation manually

The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e. accuracy) is 80% ?

Trans_id	Itemlist
T1	{K, A, D, B}
T2	{D, A C, E, B}
T3	{C, A, B, E}
T4	{B, A, D}

Hint: Make a tabular and binary representation of the data in order to better see the relationship between Items. First generate all item sets with minimum support of 60%. Then form rules and calculate their confidence base on the conditional probability P(B|A) = |B∩A| / |A|. Remember to only take the item sets from the previous phase whose support is 60% or more.

Answer 1:

Expt-5 (Apriori algorithm)

Ex-1]

Trans-id	Itemlist
T1	{K, A, D, B}
T2	{D, A, C, E, B}
T3	{C, A, B, E}
T4	{B, A, D}

(min_sup = 0.6
min_conf = 0.8)

Step 1:

Item	Support count	support
A	4	1
B	4	1
C	2	0.5
D	3	0.75
E	2	0.5
K	1	0.25

Frequent items ≡ {A}, {B}, {D}

Step 2:

Itemset	support
{A, B}	1
{A, D}	0.75
{B, D}	0.75

Frequent itemsets ≡ {A, B}, {A, D}, {B, D}

Step 3:

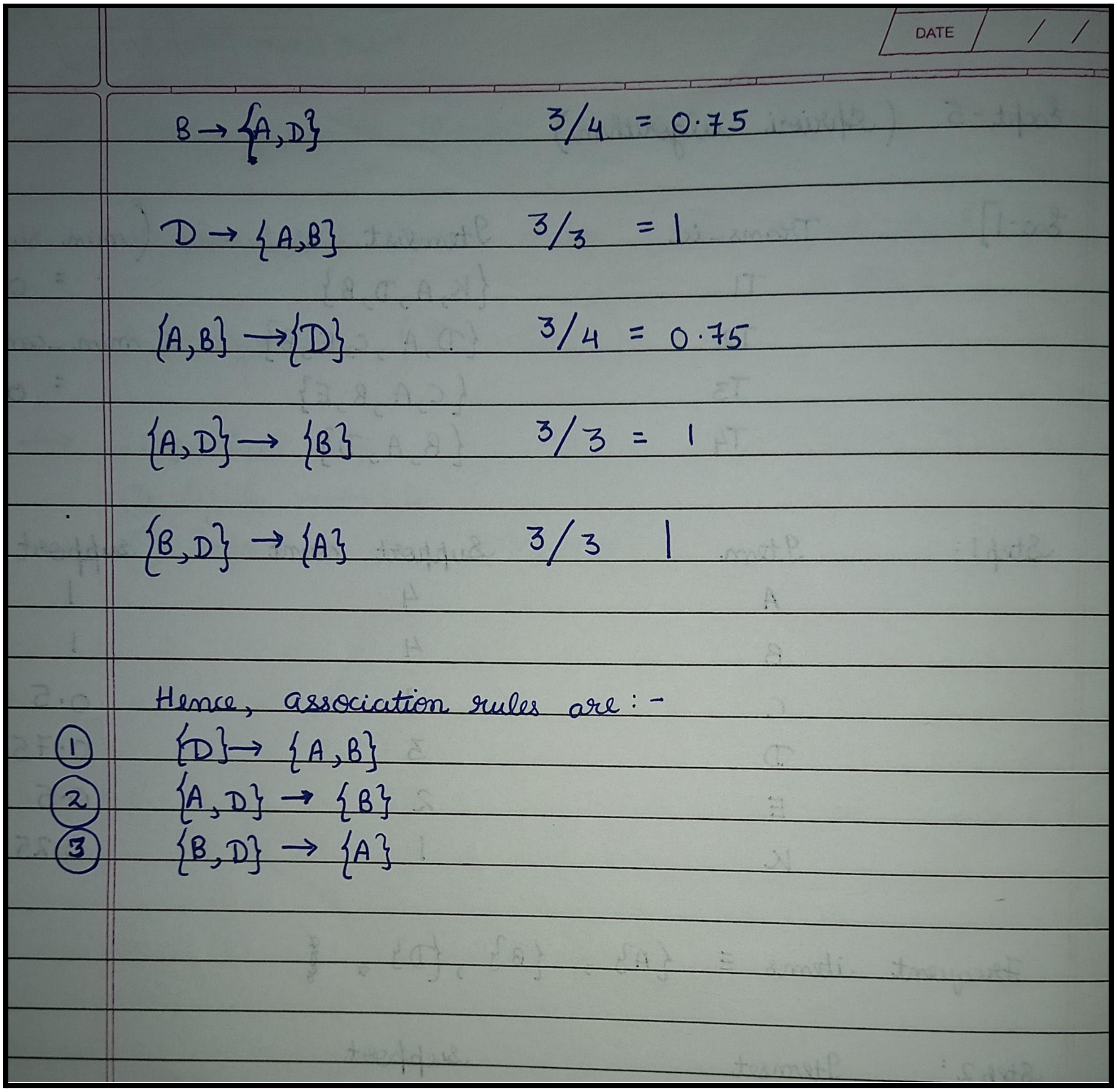
Itemset	Support
{A, B, D}	0.75

Frequent itemset ≡ {A, B, D}

→ Association rule

	confidence
A → {B, D}	$\frac{3}{4} = 0.75$

Inference: Frequent itemset calculated manually is {A,B,D} which means that support of {A,B,D} is more than threshold that is minimum support = 0.6



DATE / /

B → {A,D}	3/4 = 0.75
D → {A,B}	3/3 = 1
{A,B} → {D}	3/4 = 0.75
{A,D} → {B}	3/3 = 1
{B,D} → {A}	3/3 1

Hence, association rules are :-

1. {D} → {A,B}
2. {A,D} → {B}
3. {B,D} → {A}

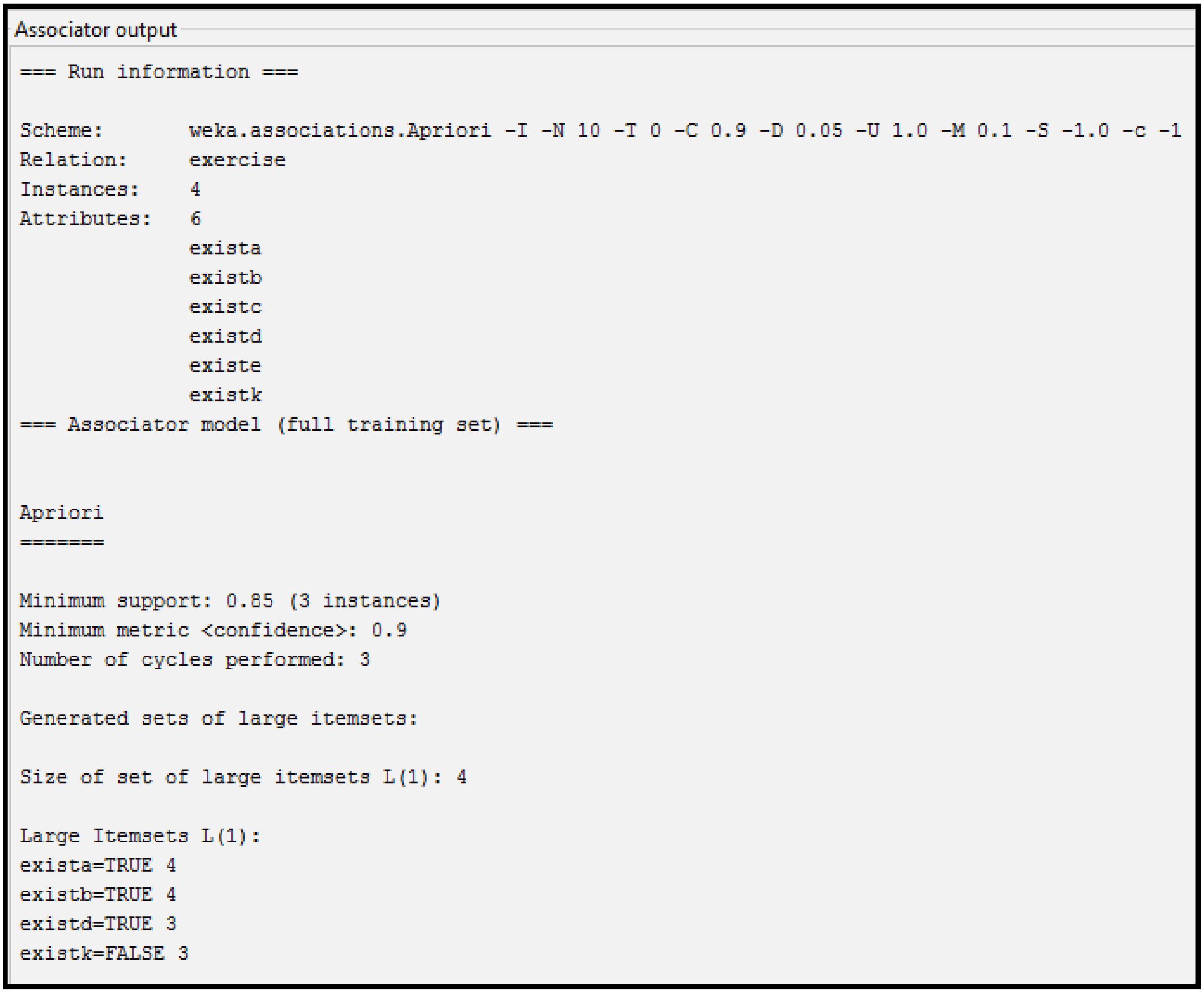
Inference:

Association rule generated manually are as shown above.These are considered because their confidence is more than threshold that is minimum confidence = 0.8

Exercise 2: Input file generation and Initial experiments with Weka's association rule discovery.

1. Launch Weka and try to do the calculations you performed manually in the previous exercise.Use the apriori algorithm for generating the association rules.

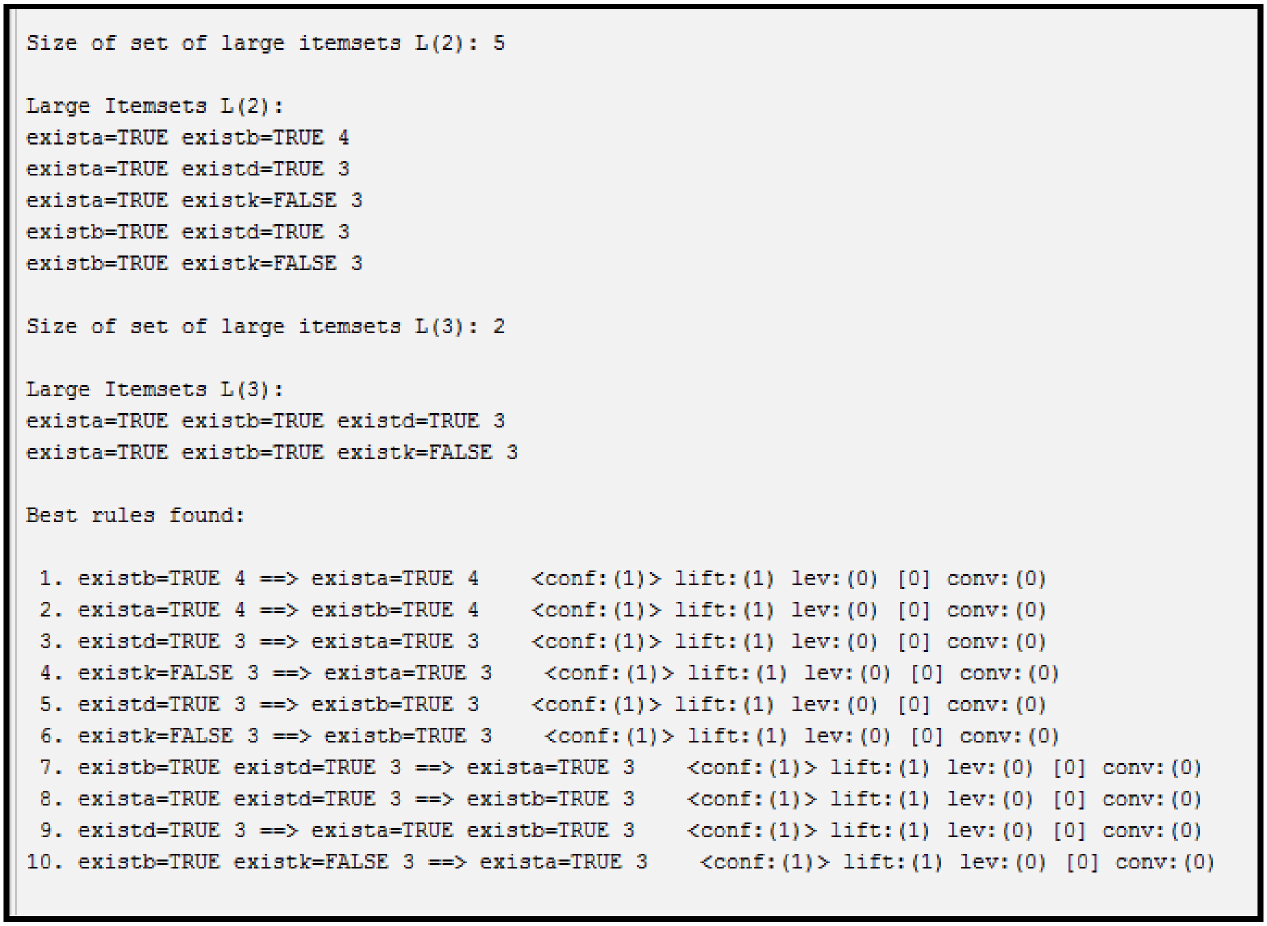
Answer 2:



```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     exercise
Instances:    4
Attributes:   6
              exista
              existb
              existc
              existd
              existe
              existk
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.85 (3 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 4

Large Itemsets L(1):
exista=TRUE 4
existb=TRUE 4
existd=TRUE 3
existk=FALSE 3
```

Inference:

Size of Large (Frequent) itemsets **containing 1 item** is 4 and it consists of {A},{B},{D},{K}.

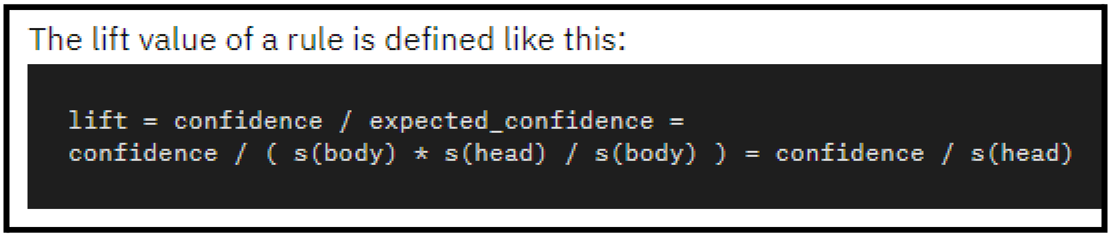


```
Size of set of large itemsets L(2): 5

Large Itemsets L(2):
exista=TRUE existb=TRUE 4
exista=TRUE existd=TRUE 3
exista=TRUE existk=FALSE 3
existb=TRUE existd=TRUE 3
existb=TRUE existk=FALSE 3

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

Large Itemsets L(3):
exista=TRUE existb=TRUE existd=TRUE 3
exista=TRUE existb=TRUE existk=FALSE 3

Best rules found:

 1. existb=TRUE 4 ==> exista=TRUE 4    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. exista=TRUE 4 ==> existb=TRUE 4    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. existd=TRUE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. existk=FALSE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. existd=TRUE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. existk=FALSE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 7. existb=TRUE existd=TRUE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 8. exista=TRUE existd=TRUE 3 ==> existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 9. existd=TRUE 3 ==> exista=TRUE existb=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. existb=TRUE existk=FALSE 3 ==> exista=TRUE 3    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

Inference:

Size of Large (Frequent) itemsets **containing 2 items** is 5 and it consists of {A,B},{A,D},{A,K},{B,D},{B,K}

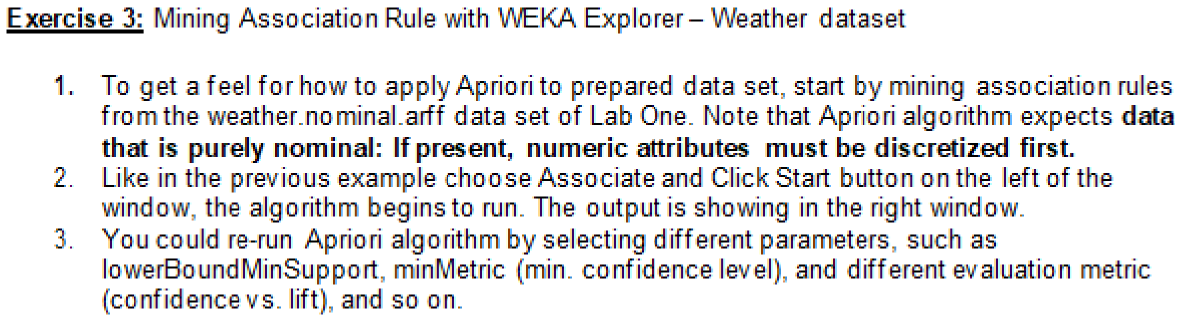
Size of Large (Frequent) itemsets **containing 3 items** is 2 and it consists of {A,B,D},{A,B,K}



The lift value of a rule is defined like this:

```
lift = confidence / expected_confidence =
confidence / ( s(body) * s(head) / s(body) ) = confidence / s(head)
```

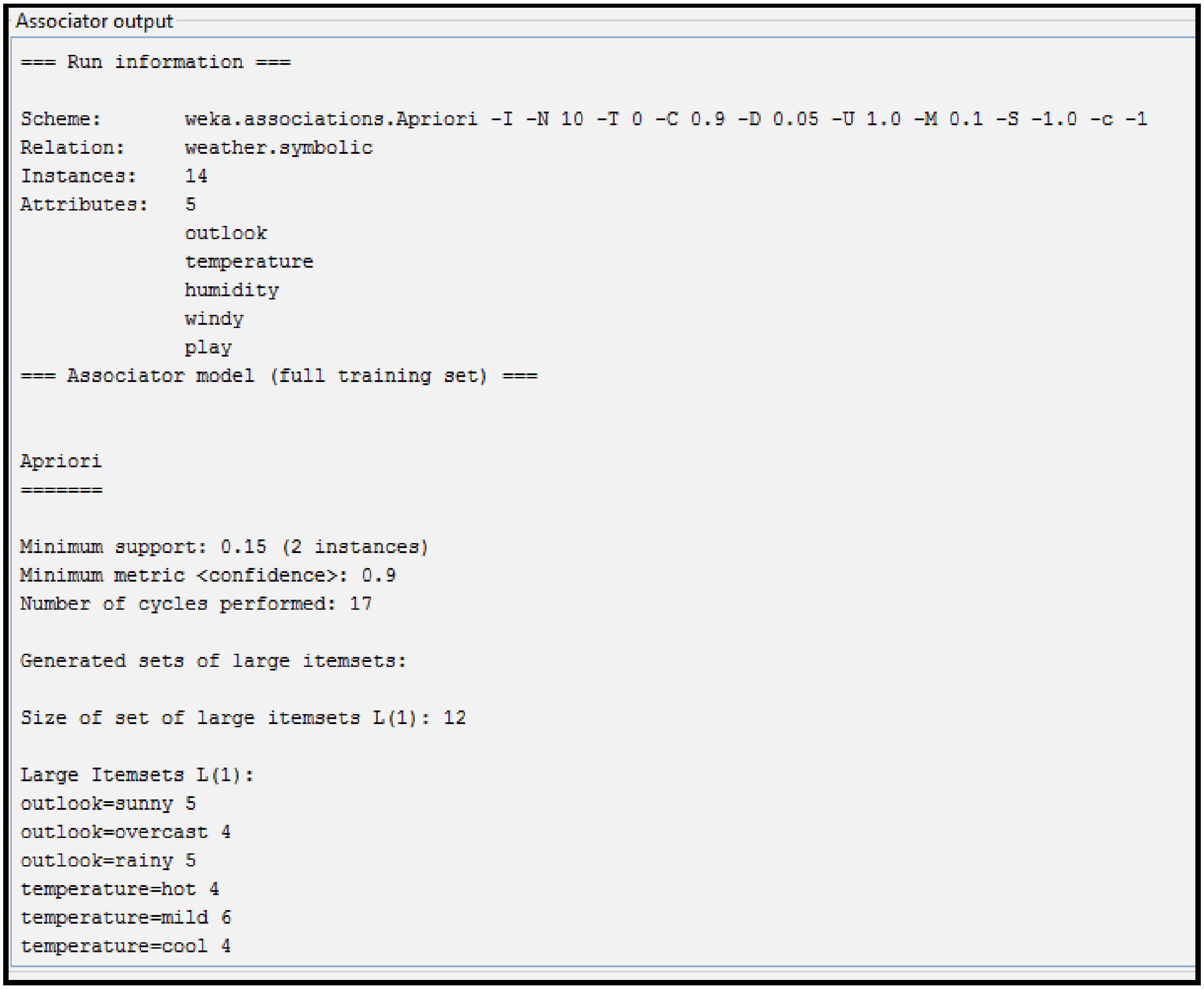
In the above, 10 best rules are generated in which we can observe that lift value is near 1 which indicates that the rule body and the rule head appear almost as often together as expected, this means that **the occurrence of the rule body has almost no effect on the occurrence of the rule head.**



Exercise 3: Mining Association Rule with WEKA Explorer – Weather dataset

1. To get a feel for how to apply Apriori to prepared data set, start by mining association rules from the weather.nominal.arff data set of Lab One. Note that Apriori algorithm expects **data that is purely nominal: If present, numeric attributes must be discretized first.**
2. Like in the previous example choose Associate and Click Start button on the left of the window, the algorithm begins to run. The output is showing in the right window.
3. You could re-run Apriori algorithm by selecting different parameters, such as lowerBoundMinSupport, minMetric (min. confidence level), and different evaluation metric (confidence vs. lift), and so on.

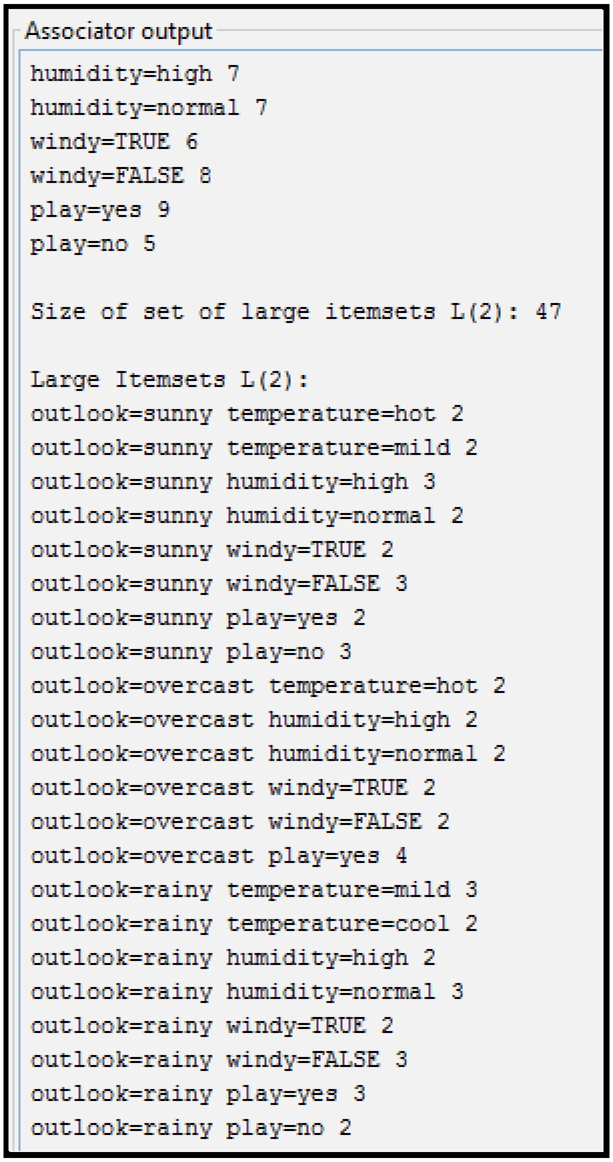
Answer 3:



```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15 (2 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Large Itemsets L(1):
outlook=sunny 5
outlook=overcast 4
outlook=rainy 5
temperature=hot 4
temperature=mild 6
temperature=cool 4
```

Inference:

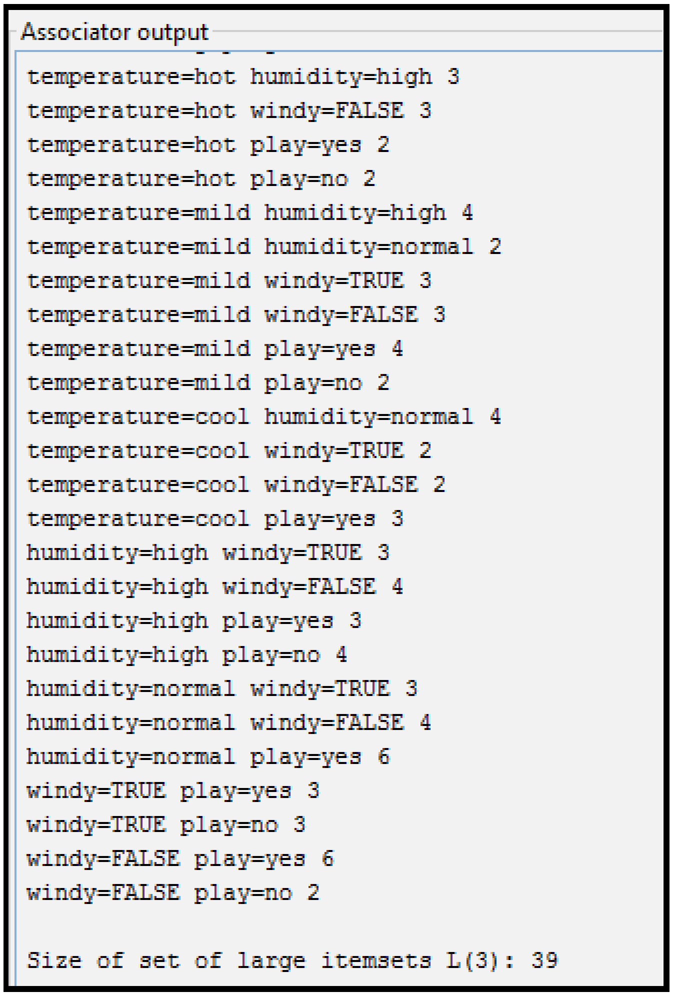
Size of Large (Frequent) itemsets **containing 1 item** is 12 and it consists of {sunny},{overcast},{rainy},{hot},etc.



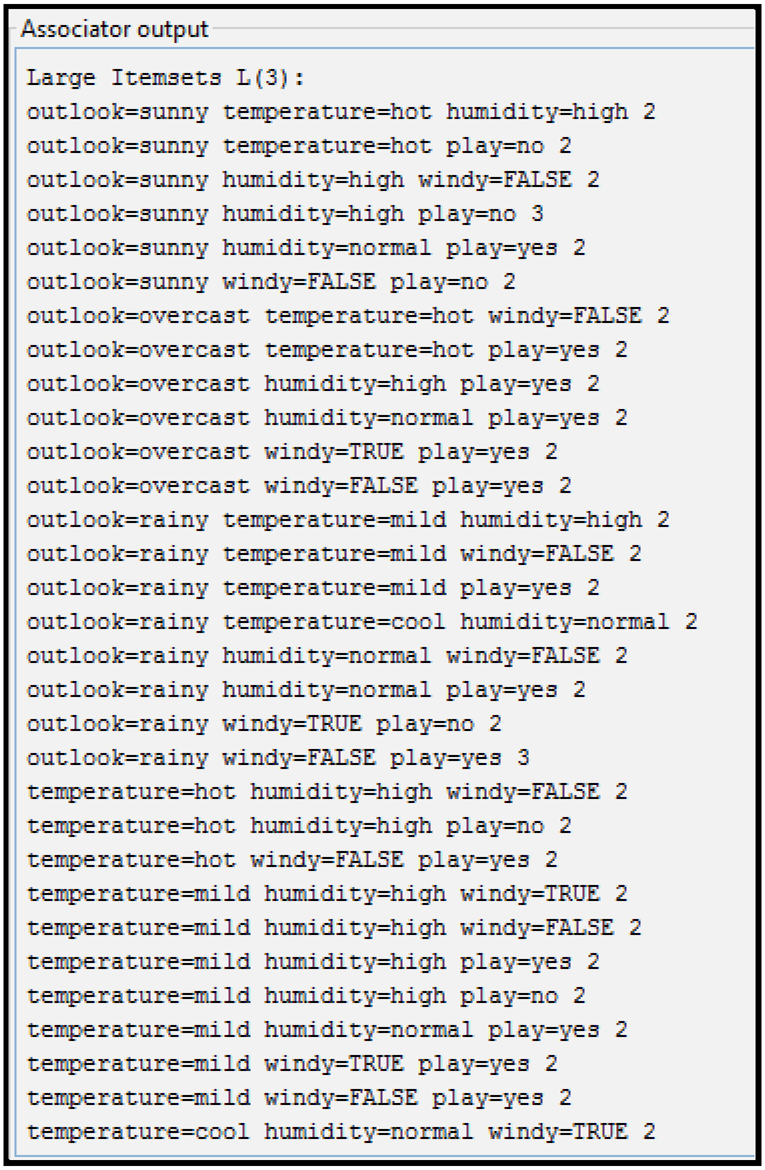
```
humidity=high 7
humidity=normal 7
windy=TRUE 6
windy=FALSE 8
play=yes 9
play=no 5

Size of set of large itemsets L(2): 47

Large Itemsets L(2):
outlook=sunny temperature=hot 2
outlook=sunny temperature=mild 2
outlook=sunny humidity=high 3
outlook=sunny humidity=normal 2
outlook=sunny windy=TRUE 2
outlook=sunny windy=FALSE 3
outlook=sunny play=yes 2
outlook=sunny play=no 3
outlook=overcast temperature=hot 2
outlook=overcast humidity=high 2
outlook=overcast humidity=normal 2
outlook=overcast windy=TRUE 2
outlook=overcast windy=FALSE 2
outlook=overcast play=yes 4
outlook=rainy temperature=mild 3
outlook=rainy temperature=cool 2
outlook=rainy humidity=high 2
outlook=rainy humidity=normal 3
outlook=rainy windy=TRUE 2
outlook=rainy windy=FALSE 3
outlook=rainy play=yes 3
outlook=rainy play=no 2
```

Inference:

Size of Large (Frequent) itemsets **containing 2 items** is 47 and it consists of {sunny,hot},{sunny,mild},{sunny,high},etc

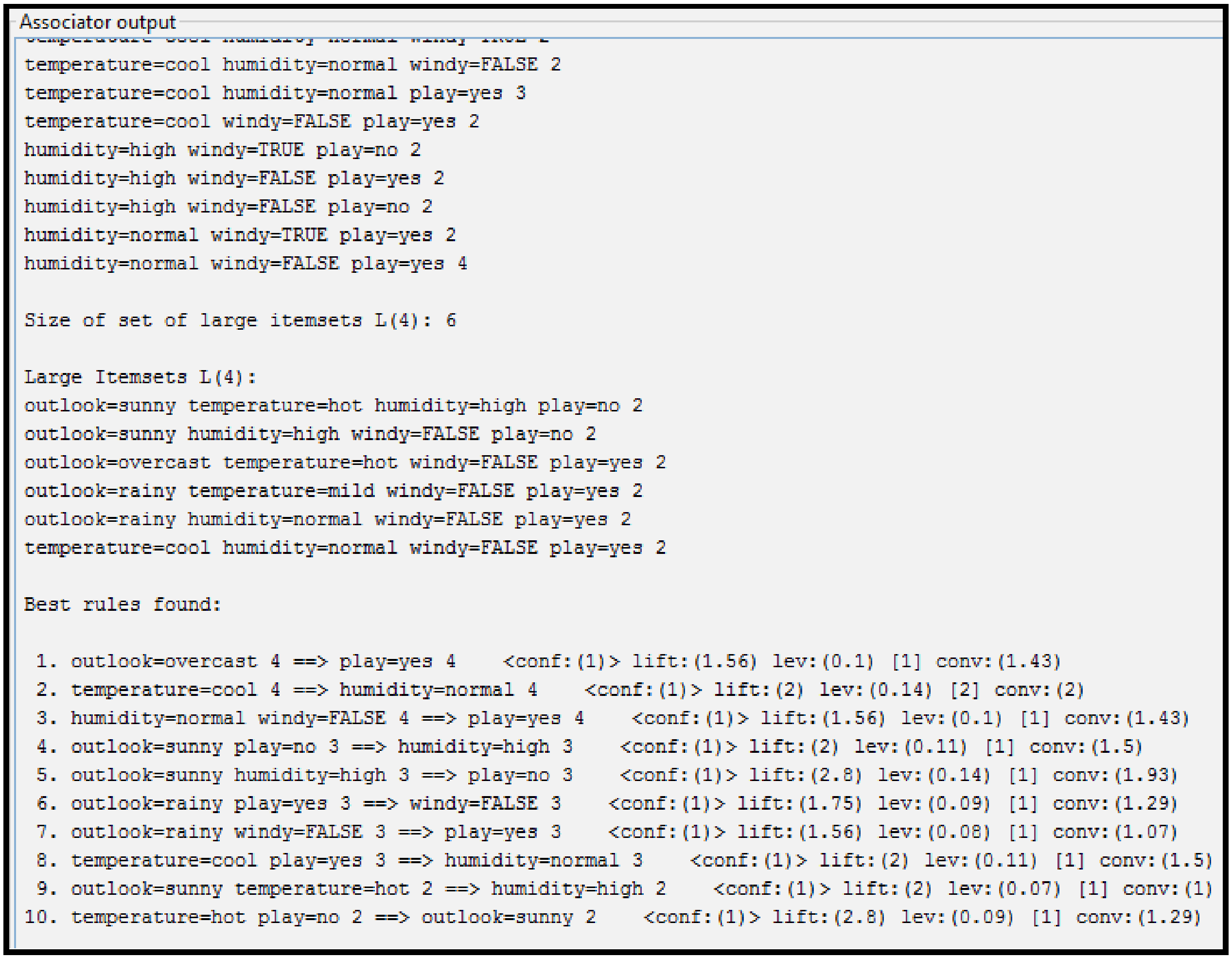


```
temperature=hot humidity=high 3
temperature=hot windy=FALSE 3
temperature=hot play=yes 2
temperature=hot play=no 2
temperature=mild humidity=high 4
temperature=mild humidity=normal 2
temperature=mild windy=TRUE 3
temperature=mild windy=FALSE 3
temperature=mild play=yes 4
temperature=mild play=no 2
temperature=cool humidity=normal 4
temperature=cool windy=TRUE 2
temperature=cool windy=FALSE 2
temperature=cool play=yes 3
humidity=high windy=TRUE 3
humidity=high windy=FALSE 4
humidity=high play=yes 3
humidity=high play=no 4
humidity=normal windy=TRUE 3
humidity=normal windy=FALSE 4
humidity=normal play=yes 6
windy=TRUE play=yes 3
windy=TRUE play=no 3
windy=FALSE play=yes 6
windy=FALSE play=no 2

Size of set of large itemsets L(3): 39
```



```
Large Itemsets L(3):
outlook=sunny temperature=hot humidity=high 2
outlook=sunny temperature=hot play=no 2
outlook=sunny humidity=high windy=FALSE 2
outlook=sunny humidity=high play=no 3
outlook=sunny humidity=normal play=yes 2
outlook=sunny windy=FALSE play=no 2
outlook=overcast temperature=hot windy=FALSE 2
outlook=overcast temperature=hot play=yes 2
outlook=overcast humidity=high play=yes 2
outlook=overcast humidity=normal play=yes 2
outlook=overcast windy=TRUE play=yes 2
outlook=overcast windy=FALSE play=yes 2
outlook=rainy temperature=mild humidity=high 2
outlook=rainy temperature=mild windy=FALSE 2
outlook=rainy temperature=mild play=yes 2
outlook=rainy temperature=cool humidity=normal 2
outlook=rainy humidity=normal windy=FALSE 2
outlook=rainy humidity=normal play=yes 2
outlook=rainy windy=TRUE play=no 2
outlook=rainy windy=FALSE play=yes 3
temperature=hot humidity=high windy=FALSE 2
temperature=hot humidity=high play=no 2
temperature=hot windy=FALSE play=yes 2
temperature=mild humidity=high windy=TRUE 2
temperature=mild humidity=high windy=FALSE 2
temperature=mild humidity=high play=yes 2
temperature=mild humidity=high play=no 2
temperature=mild humidity=normal play=yes 2
temperature=mild windy=TRUE play=yes 2
temperature=mild windy=FALSE play=yes 2
temperature=cool humidity=normal windy=TRUE 2
```

Size of Large (Frequent) itemsets **containing 3 items** is 39 and one of the example is {sunny,hot,high}

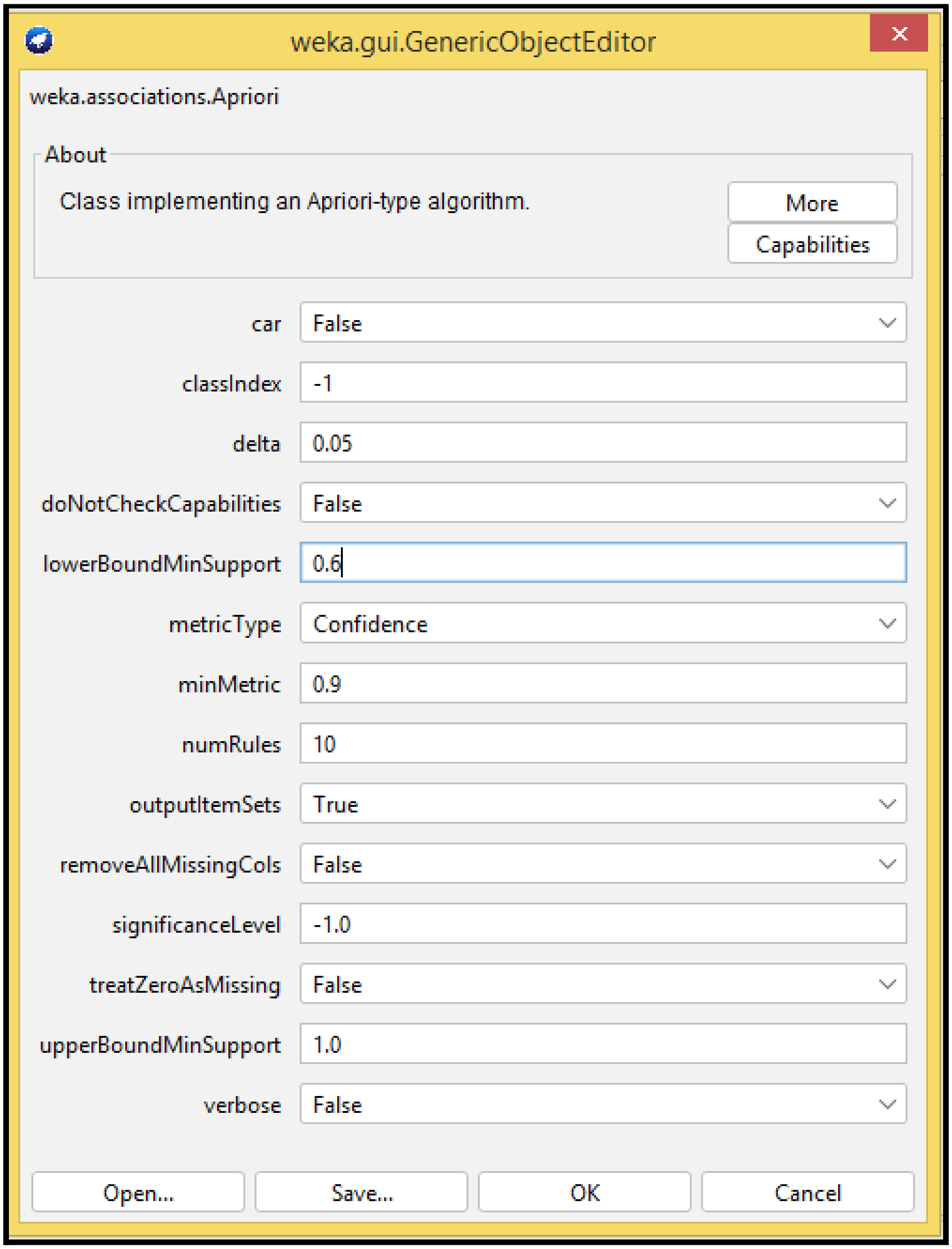


```
temperature=cool humidity=normal windy=TRUE 2
temperature=cool humidity=normal windy=FALSE 2
temperature=cool humidity=normal play=yes 3
temperature=cool windy=FALSE play=yes 2
humidity=high windy=TRUE play=no 2
humidity=high windy=FALSE play=yes 2
humidity=high windy=FALSE play=no 2
humidity=normal 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)
```

In the above, 10 best rules are generated in which we can observe that lift value is near 1 for some rules and for some greater than 1

1. A lift value greater than 1 indicates that the rule body and the rule head appear more often together than expected, this means that the occurrence of the rule body has a positive effect on the occurrence of the rule head.
2. A lift value near 1 indicates that the rule body and the rule head appear almost as often together as expected, this means that the occurrence of the rule body has almost no effect on the occurrence of the rule head.

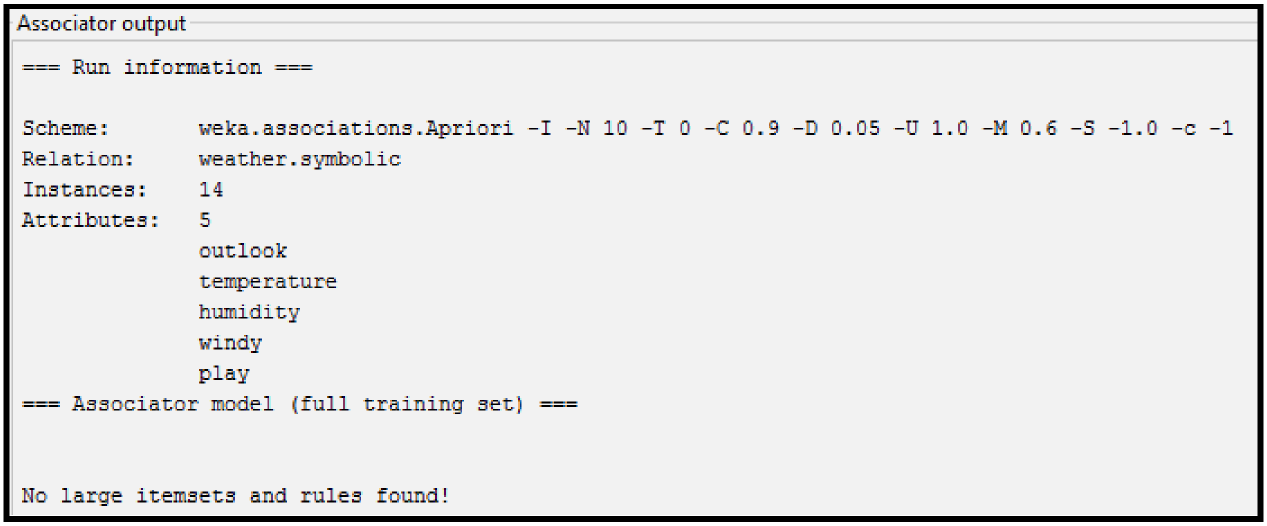
Changing lowerBoundminsupport to 0.6



weka.associations.Apriori

Class implementing an Apriori-type algorithm.

car	False
classIndex	-1
delta	0.05
doNotCheckCapabilities	False
lowerBoundMinSupport	0.6
metricType	Confidence
minMetric	0.9
numRules	10
outputItemSets	True
removeAllMissingCols	False
significanceLevel	-1.0
treatZeroAsMissing	False
upperBoundMinSupport	1.0
verbose	False



Associator output

```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.6 -S -1.0 -c -1
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
=== Associator model (full training set) ===


No large itemsets and rules found!
```

Changing lowerBoundminsupport to 0.6 does not generate any frequent itemsets which means that no itemset is having support above 0.6

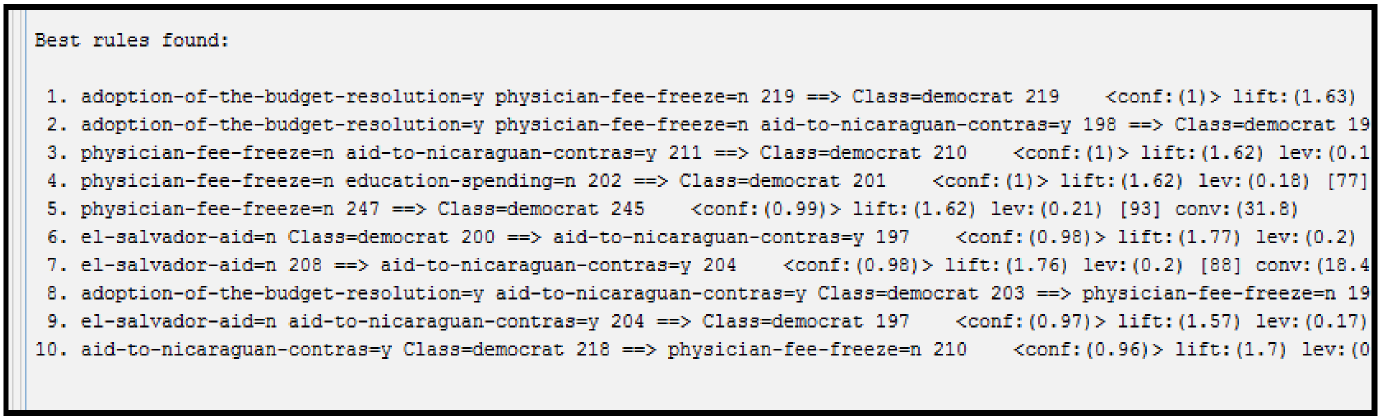
Exercise 4: Mining Association Rule with WEKA Explorer – Vote

Now consider a real-world dataset, **vote.arff**, which gives the votes of 435 U.S. congressmen on 16 key issues gathered in the mid-1980s, and also includes their party affiliation as a binary attribute. Association-rule mining can also be applied to this data to seek interesting associations.

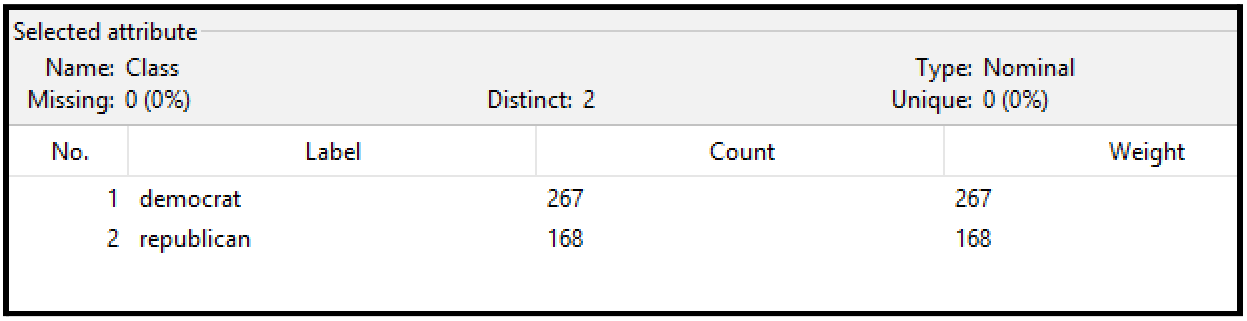
Load data at Preprocess tab. Click the Open file button to bring up a standard dialog through which you can select a file. Choose the **vote.arff** file. To see the original dataset, click the **Edit** button, a viewer window opens with dataset loaded. This is a purely nominal dataset with some missing values (corresponding to abstentions).

Task 1. Run Apriori on this data with default settings. Comment on the rules that are generated. Several of them are quite similar. How are their support and confidence values related?

Task 2. It is interesting to see that none of the rules in the default output involve Class = republican. Why do you think that is?



```
Best rules found:

 1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219    <conf:(1)> lift:(1.63)
 2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 19
 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210    <conf:(1)> lift:(1.62) lev:(0.1
 4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201    <conf:(1)> lift:(1.62) lev:(0.18) [77]
 5. physician-fee-freeze=n 247 ==> Class=democrat 245    <conf:(0.99)> lift:(1.62) lev:(0.21) [93] conv:(31.8)
 6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197    <conf:(0.98)> lift:(1.77) lev:(0.2)
 7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204    <conf:(0.98)> lift:(1.76) lev:(0.2) [88] conv:(18.4
 8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 203 ==> physician-fee-freeze=n 19
 9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197    <conf:(0.97)> lift:(1.57) lev:(0.17)
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210    <conf:(0.96)> lift:(1.7) lev:(0
```



Selected attribute

Name: Class Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	democrat	267	267
2	republican	168	168

None of the above rules include class republican because its support is less than the threshold value.

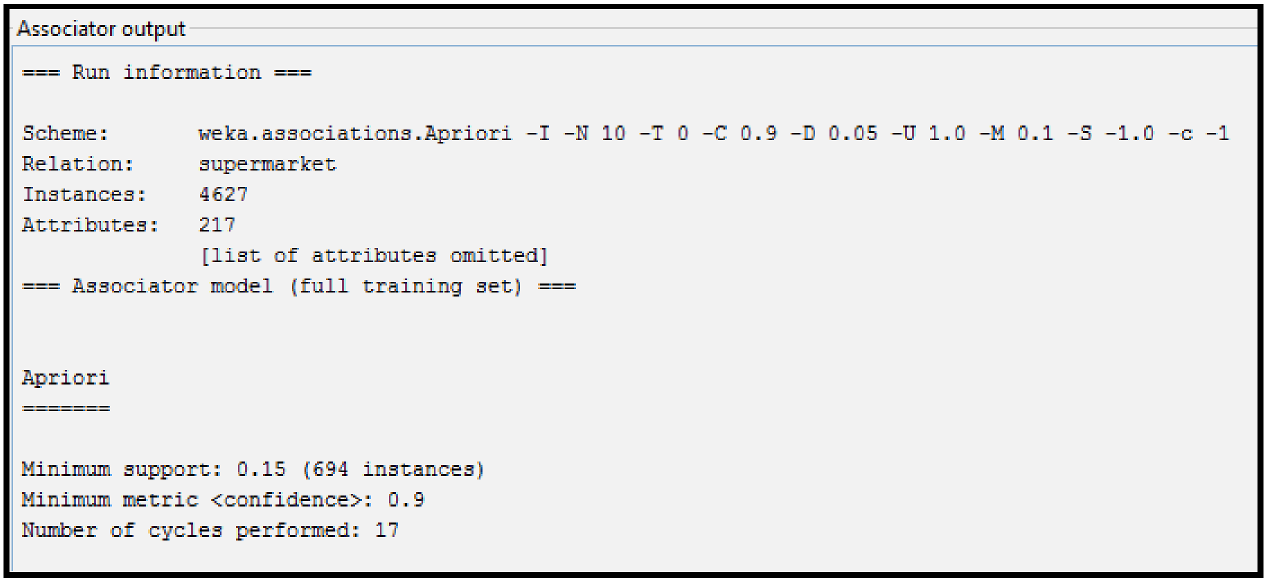
Exercise 5: Let's run Apriori on another real-world dataset.

Load data at Preprocess tab. Click the Open file button to bring up a standard dialog through which you can select a file. Choose the **supermarket.arff** file. To see the original dataset, click the **Edit** button, a viewer window opens with dataset loaded.

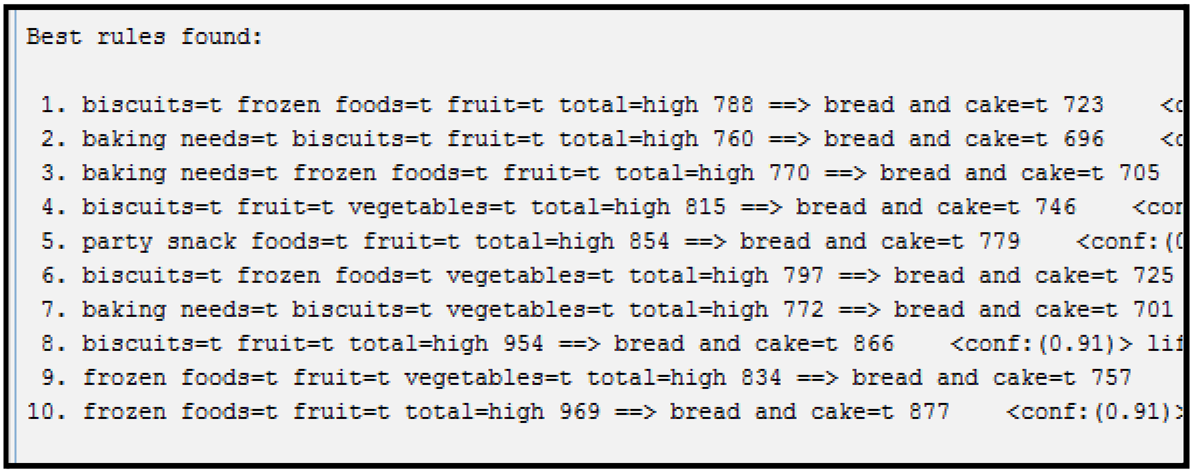
To do market basket analysis in Weka, each transaction is coded as an instance of which the attributes represent the items in the store. Each attribute has only one value: If a particular transaction does not contain it (i.e., the customer did not buy that item), this is coded as a missing value.

Task 1. Experiment with Apriori and investigate the effect of the various parameters described before. Prepare a brief oral presentation on the main findings of your investigation.

Answer 5:

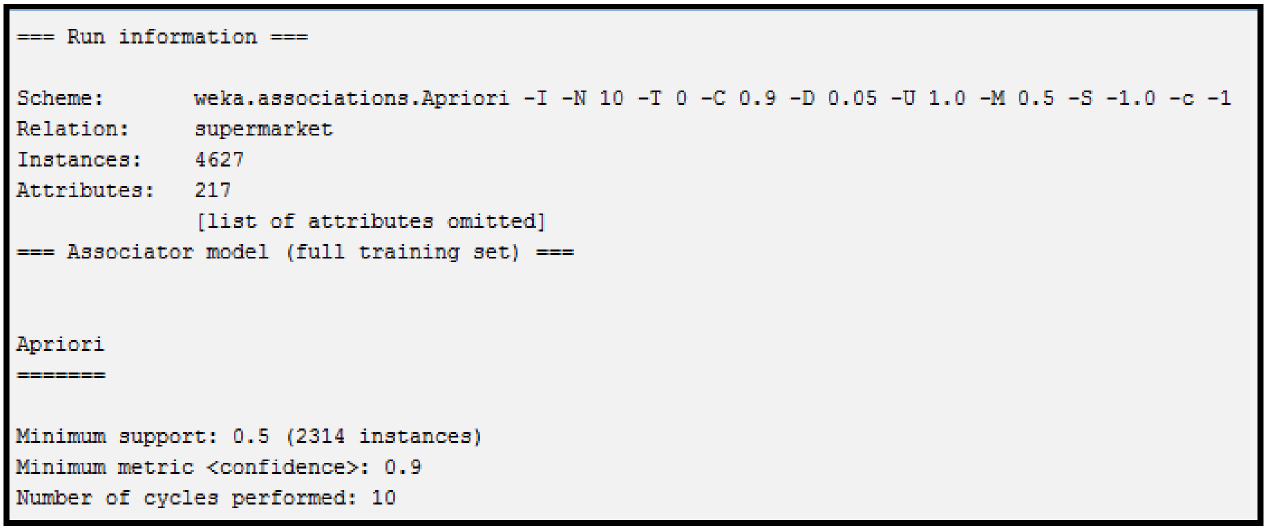


```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
              [list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17
```



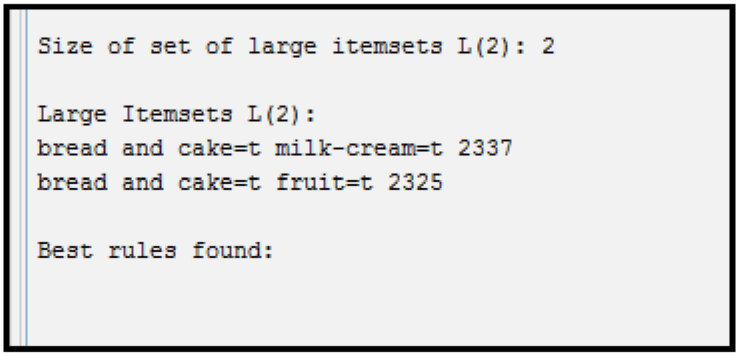
```
Best rules found:

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

Changing lowerBoundsupport to 0.5



```
=== Run information ===

Scheme:       weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.5 -S -1.0 -c -1
Relation:     supermarket
Instances:    4627
Attributes:   217
              [list of attributes omitted]
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.5 (2314 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10
```

No best rules are found



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

Large Itemsets L(2):
bread and cake=t milk-cream=t 2337
bread and cake=t fruit=t 2325

Best rules found:
```

Similarly no best rules are generated for lowerBoundsupport equal to 0.4 or 0.3 or 0.2

Conclusion:

Apriori algorithm uses a generate and test approach that is generates candidate itemsets and tests if they are frequent.It is breadth first search and terminates when no frequent or candidate set can be generated.We also observed that if an itemset is frequent then all of its subsets must also frequent.Here we performed Apriori algorithm on WEKA and observed that if we change parameters such as lower bound minimum support then large itemsets change.We also observed that sometimes no large itemsets is generated because lower bound minimum support was too high.