Week 2 and Week 3 Tutorial exercises (March 17th and 24th 2006)
Association Rule Mining.

Exercise 1. ***Apriori***

Trace the results of using the Apriori algorithm on the grocery store example with support threshold *s*=33.34% and confidence threshold *c*=60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

Solution:

Support threshold =33.34% => threshold is at least 2 transactions.

Applying Apriori

Pass (k)	Candidate k-itemsets and their support	Frequent k-itemsets
k=1	HotDogs(4), Buns(2), Ketchup(2), Coke(3), Chips(4)	HotDogs, Buns, Ketchup, Coke, Chips
k=2	{HotDogs, Buns}(2), ~~{HotDogs, Ketchup}~~(1), {HotDogs, Coke}(2), {HotDogs, Chips}(2), ~~{Buns, Ketchup}~~(1), ~~{Buns, Coke}~~(0), ~~{Buns, Chips}~~(0), ~~{Ketchup, Coke}~~(0), ~~{Ketchup, Chips}~~(1), {Coke, Chips}(3)	{HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}
k=3	{HotDogs, Coke, Chips}(2)	{HotDogs, Coke, Chips}
k=4	{}	

Note that {HotDogs, Buns, Coke} *and* {HotDogs, Buns, Chips} *are not candidates when* k=3 *because their subsets* {Buns, Coke} *and* {Buns, Chips} *are not frequent*.

Note also that normally, there is no need to go to k=4 *since the longest transaction has only 3 items.*

All Frequent Itemsets: {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips}, {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}, {HotDogs, Coke, Chips}.

Association rules:

{HotDogs, Buns} would generate: HotDogs → Buns (2/6=0.33, 2/4=0.5) and
Buns → HotDogs (2/6=0.33, 2/2=1);

{HotDogs, Coke} would generate: HotDogs → Coke (0.33, 0.5) and
Coke → HotDogs (2/6=0.33, 2/3=0.66);

{HotDogs, Chips} would generate: HotDogs → Chips (0.33, 0.5) and
Chips → HotDogs (2/6=0.33, 2/4=0.5);

{Coke, Chips} would generate: **Coke → Chips (3/6=0.5, 3/3=1)** and
Chips → Coke (3/6=0.5, 3/4=0.75);

{HotDogs, Coke, Chips} would generate: HotDogs → Coke ^ Chips (2/6=0.33, 2/4=0.5),
Coke → Chips ^ HotDogs (2/6=0.33, 2/3=0.66),
Chips → Coke ^ HotDogs (2/6=0.33, 2/4=0.5),
HotDogs ^ Coke → Chips(2/6=0.33, 2/2=1),
HotDogs ^ Chips → Coke(2/6=0.33, 2/2=1) and
Coke ^ Chips → HotDogs(2/6=0.33, 2/3=0.66).

With the confidence threshold set to 60%, the Strong Association Rules are (sorted by confidence):

1. Coke → Chips (0.5, 1)
2. Buns → HotDogs (0.33, 1);
3. HotDogs ^ Coke → Chips(0.33, 1)
4. HotDogs ^ Chips → Coke(0.33, 1)
5. Chips → Coke (0.5, 0.75);
6. Coke → HotDogs (0.33, 0.66);
7. Coke → Chips ^ HotDogs (0.33, 0.66)
8. Coke ^ Chips → HotDogs(0.33, 0.66).

Exercise 2. ***FP-tree and FP-Growth***

a) Use the transactional database from the previous exercise with same support threshold and build a frequent pattern tree (FP-Tree). Show for each transaction how the tree evolves.
b) Use Fp-Growth to discover the frequent itemsets from this FP-tree.

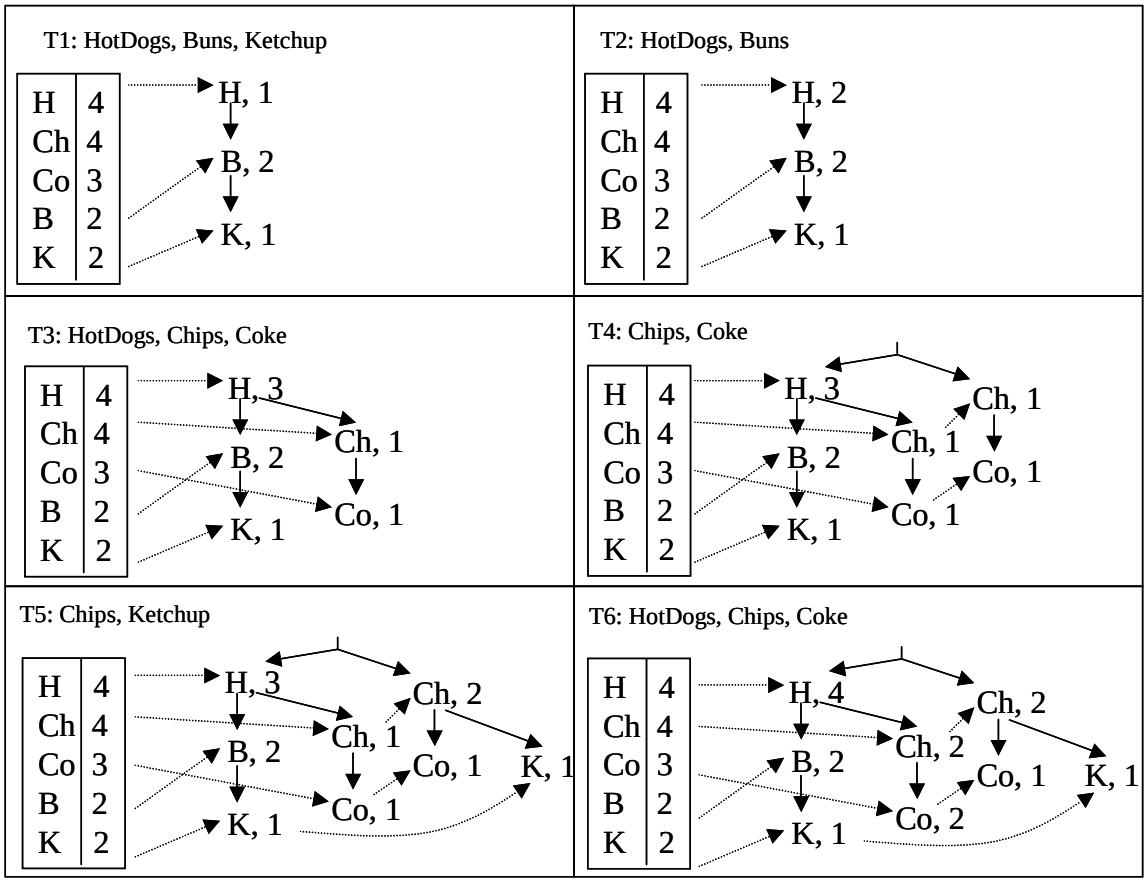
Solution:

a) The first scan of the database generates the list of frequent 1-itemsets and builds the header table where the items are sorted by frequency.

Error!

Item	Code	Support
HotDogs	H	4 = 66%
Chips	Ch	4 = 66%
Coke	Co	3 = 50%
Buns	B	2 = 33%
Ketchup	K	2 = 33%

The second scan is used to create the FP-tree. Each transaction is sorted by item support.



b) We need to build a conditional tree for each frequent item starting from the least frequent.

- For Ketchup (K), we have two branches H-B-K and Ch-K but since K has a support of 1 in each branch, this would eliminate all items (since support threshold is 2) leaving only <K:2>. This leads to the

discovery of {Ketchup} (2) as frequent item.

- For Buns (B), we have only one branch H-B. The sub-transaction {HotDogs, Buns} appears twice. We have thus the patterns <B:2, H:2> and <B:2>. This leads to the discovery of {HotDogs, Buns} (2) and {Buns}(2) as frequent itemsets.
- For Coke (Co), we have two branches: H-Ch-Co and Ch-Co resulting in the tree Co(3)→Ch(3)→H(2). We have thus 3 patterns: <Co:2, Ch:2, H:2>, <Co:3, Ch:3> and <Co:3>. This leads to the discovery of the following frequent itemsets: {Coke, Chips, HotDogs}(2), {Coke, Chips}(3) and {Coke}(3).
- For Chips (Ch), we have two paths H-Ch and Ch, giving the following tree Ch(4)→H(2). This gives the patterns <Ch:2, H:2> and <Ch:4>. Thus, the itemsets {Chips, HotDogs}(2) and {Chips}(4) are frequent.
- For HotDogs (H), The only and obvious pattern is <H:4> leading to the discovery of {HotDogs}(4) as frequent itemset.

All Frequent Itemsets (like in previous exercise): {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips}, {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}, {HotDogs, Coke, Chips}.

Notice that there was no candidacy generation. Frequent itemsets were generated directly.

Exercise 3: Using WEKA

Load a dataset described with nominal attributes, e.g. weather.nominal. Run the Apriori algorithm to generate association rules.

Solution:

Running Weka with the default parameters:
Apriori –N 10 –T 0 –C 0.9 –D 0.05 –U 1.0 –M 0.1 –S -1.0

```
=== Run information ===

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

Apriori
=======

Minimum support: 0.15
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

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

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

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

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

Best rules found:

 1. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
 2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
 3. outlook=overcast 4 ==> play=yes 4    conf:(1)
```

```
 4. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
 5. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
 6. outlook=rainy play=yes 3 ==> windy=FALSE 3    conf:(1)
 7. outlook=sunny humidity=high 3 ==> play=no 3    conf:(1)
 8. outlook=sunny play=no 3 ==> humidity=high 3    conf:(1)
 9. temperature=cool windy=FALSE 2 ==> humidity=normal play=yes 2    conf:(1)
10. temperature=cool humidity=normal windy=FALSE 2 ==> play=yes 2    conf:(1)
```

Exercise 4: **Apriori and FP-Growth** **(to be done at your own time, not in class)**

Giving the following database with 5 transactions and a minimum support threshold of 60% and a minimum confidence threshold of 80%, find all frequent itemsets using (a) Apriori and (b) FP-Growth. (c) Compare the efficiency of both processes. (d) List all strong association rules that contain "A" in the antecedent (Constraint). (e) Can we use this constraint in the frequent itemset generation phase?

TID	Transaction
T1	{A, B, C, D, E, F}
T2	{B, C, D, E, F, G}
T3	{A, D, E, H}
T4	{A, D, F, I, J}
T5	{B, D, E, K}