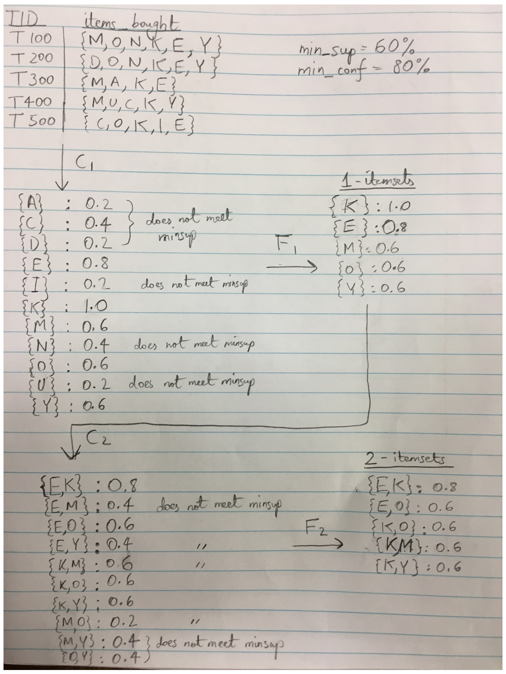
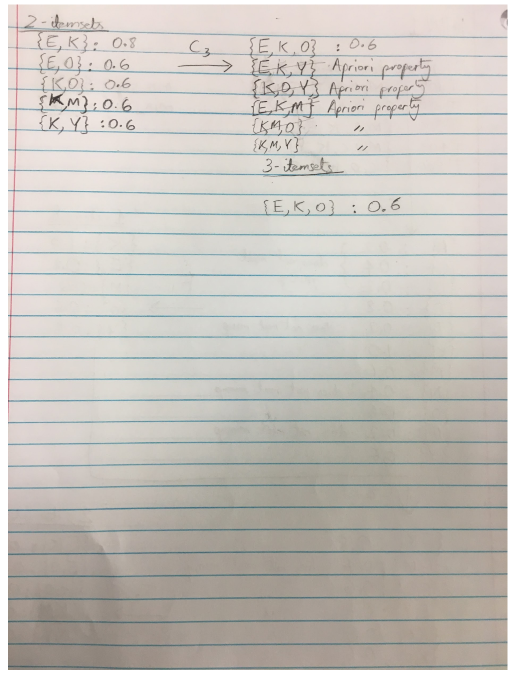
CSCI 349 Homework 1

Exercise 1





a.

Frequent Itemset	Support
{K}	1
{E}	0.8
{M}	0.6
{O}	0.6
{Y}	0.6
{E, K}	0.8
{E, O}	0.6
{K, O}	0.6
{K, M}	0.6
{K, Y}	0.6
{E, K, O}	0.6

b. A closed frequent itemset refers to a frequent itemset for which none of its immediate supersets (e.g. a 2 itemset with an additional item) has the same support as the itemset. In the example above, we have the following frequent closed itemsets:
1-itemsets (closed): {K}
2-itemsets (closed): {E, K}, {K, M}, {K, Y}
3-itemsets (closed): {E, K, O}

c. A max frequent itemset refers to a frequent itemset for which none of its immediate itemsets are frequent. In our example above, we have the following maximal frequent itemsets:

1-itemsets (closed): None
2-itemsets (closed): {K, M}, {K, Y}
3-itemsets (closed): {E, K, O}

d. **{E, K, O}** has subsets {E, K}, {E, O}, {K, O}, {E}, {K}, {O}.

Rule	Confidence	Lift
~~{E,K} → {O}~~	~~$0.6/0.8 = 0.75$~~	~~$0.6/(0.8 * 0.6) = 1.25$~~
{E,O} → {K}	$0.6/0.8 = 0.75$	$0.6/(0.8 * 0.6) = 1.25$
{K,O} → {E}	$0.6/0.8 = 0.75$	$0.6/(0.8 * 0.6) = 1.25$
~~{E} → {K,O}~~	~~$0.6/0.8 = 0.75$~~	~~$0.6/(0.8 * 0.6) = 1.25$~~
~~{K} → {E,O}~~	~~$0.6/0.8 = 0.75$~~	~~$0.6/(0.8 * 0.6) = 1.25$~~
{O} → {E,K}	$0.6/0.8 = 0.75$	$0.6/(0.8 * 0.6) = 1.25$

{K, Y} has subsets {K}, {Y}.

Rule	Confidence	Lift
~~{K} → {Y}~~	~~$0.6/1.0 = 0.6$~~	~~$0.6/(0.6 * 1.0) = 1.0$~~
{Y} → {K}	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$

{K, M} has subsets {K}, {M}.

Rule	Confidence	Lift
~~{K} → {M}~~	~~$0.6/1.0 = 0.6$~~	~~$0.6/(0.6 * 1.0) = 1.0$~~
{M} → {K}	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$

{K, O} has subsets {K}, {O}.

Rule	Confidence	Lift
~~{K} → {O}~~	~~$0.6/1.0 = 0.6$~~	~~$0.6/(0.6 * 1.0) = 1.0$~~
{O} → {K}	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$

{E,O} has subsets {E}, {O}.

Rule	Confidence	Lift
~~{E} → {O}~~	~~$0.6/0.8 = 0.75$~~	~~$0.6/(0.6 * 0.8) = 1.25$~~
{O} → {E}	$0.6/0.6 = 1.0$	$0.6/(0.6 * 0.8) = 1.25$

{E, K} has subsets {E}, {K}.

Rule	Confidence	Lift
{E} → {K}	$0.8/0.8 = 1.0$	$0.8/(0.8 * 1.0) = 1.0$
{K} → {E}	$0.8/1.0 = 0.8$	$0.8/(1.0 * 0.8) = 1.0$

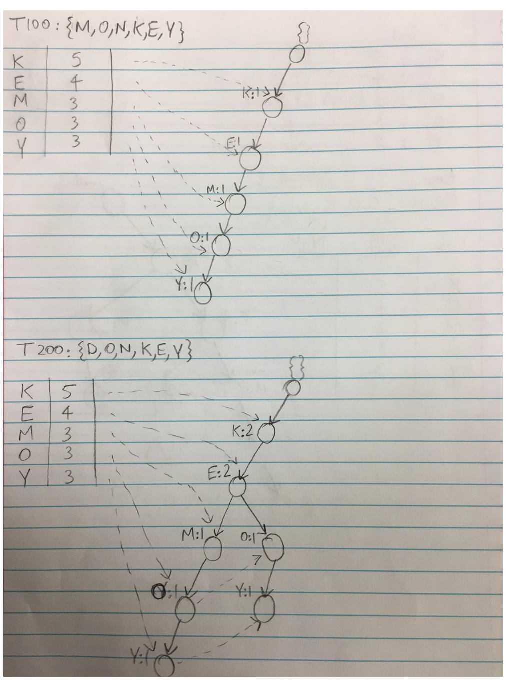
Compiled together, we have the following strong association rules:

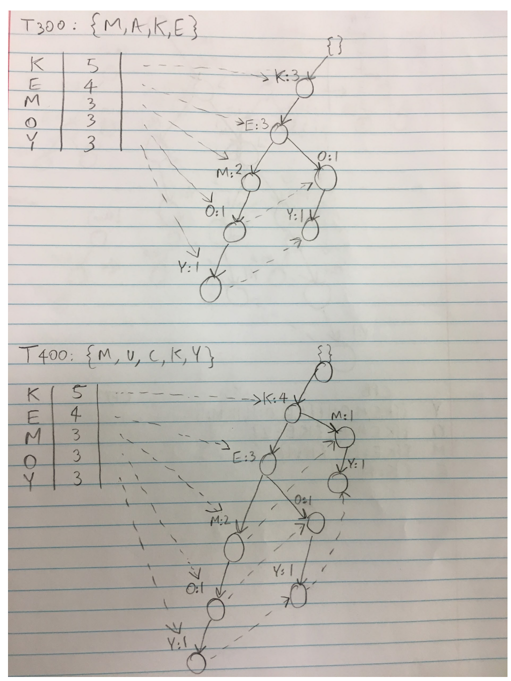
Rule	**Confidence**	**Lift**
$\{E, O\} \rightarrow \{K\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$
$\{K, O\} \rightarrow \{E\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 0.8) = 1.25$
$\{O\} \rightarrow \{E, K\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 0.8) = 1.25$
$\{Y\} \rightarrow \{K\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$
$\{M\} \rightarrow \{K\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$
$\{O\} \rightarrow \{K\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 1.0) = 1.0$
$\{O\} \rightarrow \{E\}$	$0.6/0.6 = 1.0$	$0.6/(0.6 * 0.8) = 1.25$
$\{E\} \rightarrow \{K\}$	$0.8/0.8 = 0.75$	$0.8/(0.8 * 1.0) = 1.0$
$\{K\} \rightarrow \{E\}$	$0.8/1.0 = 0.8$	$0.8/(1.0 * 0.8) = 1.0$

e. From the table above, we can observe that $\{K, O\} \rightarrow \{E\}$ and $\{O\} \rightarrow \{E, K\}$ are two of the strongest rules based on confidence (1.0), lift (1.25), and size of antecedent and consequent itemsets.

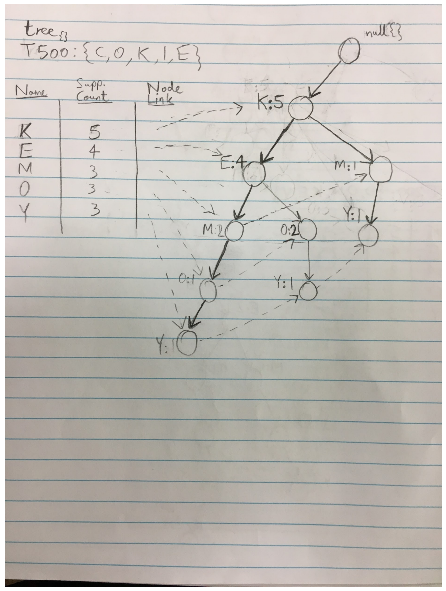
Exercise 2

a. & b. Work shown below.

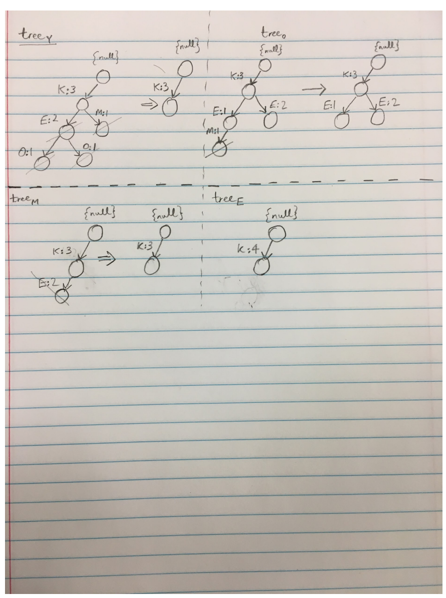




$tree_{\{\}}$



c. Frequent Patterns



Item	Conditional Pattern Base	Conditional FP-Tree	Freq. patterns
Y	{{K, E, M, O: 1}, {K, E, O: 1}, {K, M: 1}}	{K: 3}	{K, Y: 3}
O	{{K, E, M: 1}, {K, E: 2}}	{E: 3, K: 3}	{K, O: 3}, {E, O: 3}, {E, K, O: 3}
M	{{K, E: 2}, {K: 1}}	{K: 3}	{K, M: 3}
E	{{K: 4}}	{K: 4}	{E, K: 4}

d. From a computational point of view, the Apriori algorithm is more compact, but has a longer runtime whereas the FP-growth algorithm has a greater space requirement to hold all the pointers, but has a shorter runtime.

Exercise 3

a. Vertical data format

Item	TID
A	T300
C	T400, T500
D	T200
E	T100, T200, T300, T500
I	T500
K	T100, T200, T300, T400, T500
M	T100, T300, T400
N	T100, T200
O	T100, T200, T400
U	T400
Y	T100, T200, T400

b. Frequent 1-itemsets

Item	TID
E	T100, T200, T300, T500
K	T100, T200, T300, T400, T500
M	T100, T300, T400
O	T100, T200, T500
Y	T100, T200, T400

Frequent 2-itemsets

Item	TID
{E, K}	T100, T200, T300, T500
{E, O}	T100, T200, T500
{K, O}	T100, T200, T500
{K, M}	T100, T300, T400
{K, Y}	T100, T200, T400

Frequent 3-itemsets

Item	TID
{E, K, 0}	T100, T200, T500

Exercise 4

	A	NOT A	Total
B	65	40	105
NOT B	35	10	45
Total	100	50	150

a. $min\_sup = 0.4$ and $min\_conf = 0.6$

$supp(A \to B) = P(A \cup B) = \frac{65}{150} = 0.43$

$conf(A \to B) = P(B|A) = \frac{0.43}{100/150} = 0.64$

The rule is strong because it satisfies both min support and confidence thresholds

b. The lift measure tells us if two items or itemsets are correlated or independent and if they are correlated, whether it's a positive or negative correlation.

$lift(A, B) = \frac{0.43}{(100/150) \cdot (105/150)} = \frac{0.43}{0.47} = 0.92$

This suggests that A and B have a negative correlation and are not likely to occur together. Therefore this is not a good rule.

c. Expected values

	A	NOT A	Total
B	70	35	105
NOT B	30	15	45
Total	100	50	150

d. χ^2 correlation coefficient $= \frac{(70-65)^2}{70} + \frac{(35-40)^2}{35} + \frac{(30-35)^2}{30} + \frac{(15-10)^2}{15} = 3.57$. The p-value is .058782, which is not significant at $p < 0.05$ and so dependency is not implied

e. $supp(A \to \overline{B}) = P(A \cup \overline{B}) = \frac{30}{150} = 0.2$
$conf(A \to \overline{B}) = P(\overline{B}|A) = \frac{0.2}{100/150} = 0.3$
$lift(A, \overline{B}) = \frac{0.2}{(100/150)\cdot(45/150)} = 1$

f. $conf(\overline{B} \to A) = P(A|\overline{B}) = \frac{0.2}{45/150} = 0.67$
$lift(\overline{B}, A) = \frac{0.2}{(45/150)\cdot(100/150)} = 1$
Th rule $\overline{B} \to A$ is stronger because it has higher confidence.

g. $Kulc(A, \overline{B}) = \frac{1}{2} \cdot \left(\frac{45}{150} + \frac{100}{150}\right) = 0.5 \cdot (0.67 + 0.3) = 0.485$

h.

$$
\begin{aligned}
IR(A, \overline{B}) &= \frac{|P(A) - P(\overline{B})|}{P(A) + P(\overline{B}) - P(A \cup \overline{B})} \\
&= \frac{|\frac{100}{150} - \frac{45}{150}|}{\frac{100}{150} + \frac{45}{150} - \frac{30}{150}} \\
&= 0.478
\end{aligned}
$$

Exercise 5

This can be accomplished using distributed FP-growth algorithm.[1, 2]

Algorithm 1: Global Association Rules

Data: Derive global association rules for some database D
Split D into n partitions
for $j = 1$ **to** n **do**
    Find local frequent itemsets for D_j
Create candidate itemset table by combining all local frequent itemsets $D_1, D_2, \ldots, D_n$
Find global frequent itemsets from the candidates
return Frequent itemsets in D

[1] *Han, Jiawei, et al. Data Mining: Concepts and Techniques, Elsevier Science & Technology, 2011.*
[2] *Li, Haoyuan, et al. "PFP: parallel FP-growth for query recommendation." Proceedings of the 2008 ACM conference on Recommender systems. ACM, 2008.*