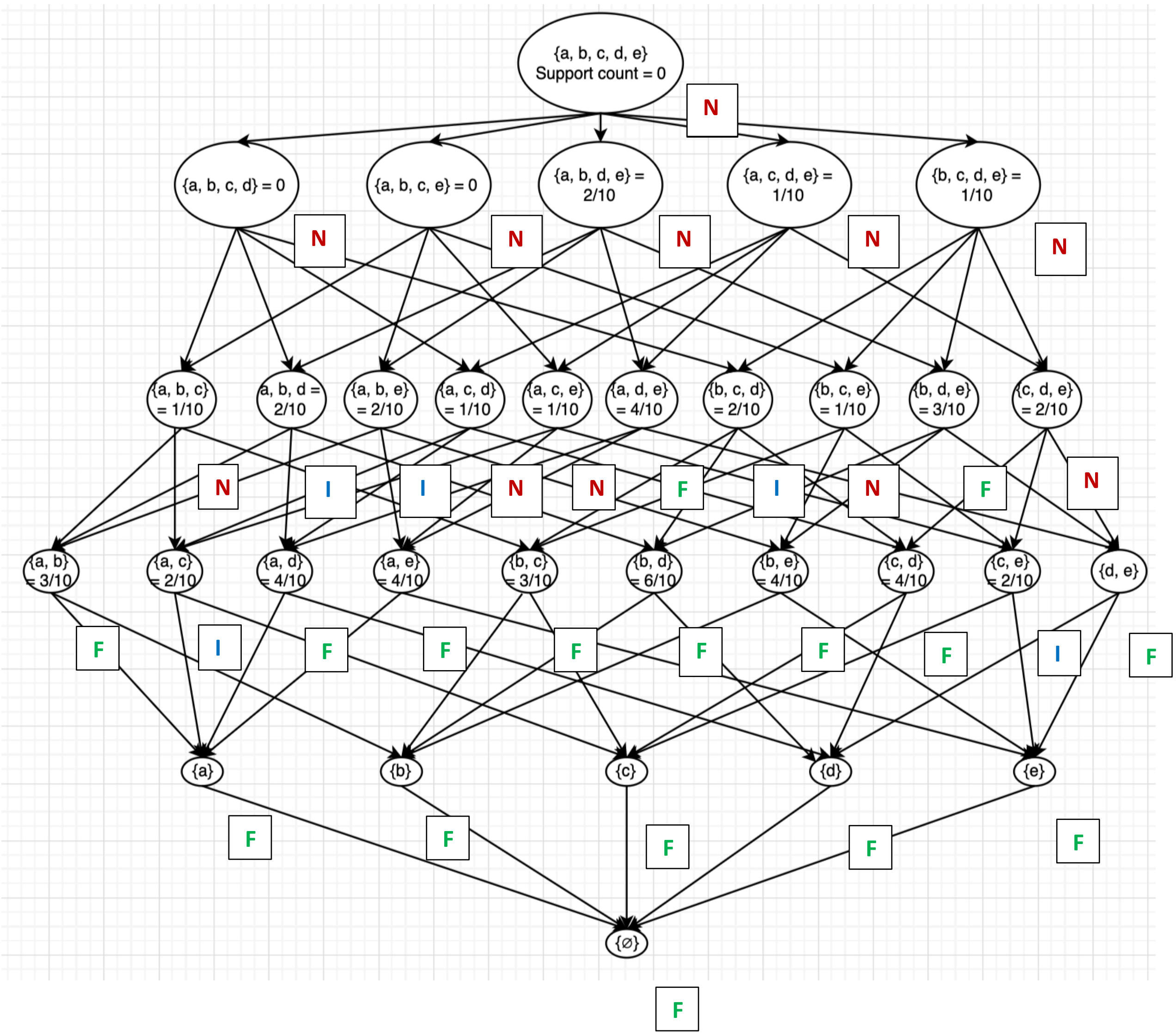
Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

1.

(a) Draw an itemset lattice representing the data



(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

There are 32 itemsets. There are 16 itemsets that are considered *frequent* as they have a support level of greater than or equal to 30%.

Thus, the percentage of frequent itemsets is 16/32 = 50%

(c) What is the pruning ratio of the Apriori algorithm on this data set?

The pruning ratio is the total number of “N” itemsets in which the itemset is not considered to be a candidate item set by the Apriori algorithm divided by the total number of itemsets. This would be 11/32 = **34.38%**

(d) What is the false alarm rate (i.e., percentage of candidate itemsets that are found to be infrequent after performing support counting?)

The false alarm rate is the total number of “I” itemsets in which the itemsets that were found to be infrequent after support counting, divided by the total number of itemsets. This would be 4/32 = **12.5%.**

2. Consider the `yelp5.csv` data

How many frequent itemsets are possible?

Among the 11 discrete attributes, each attribute has a number of unique values it can take on. For example, some only take on the binary “TRUE” or “FALSE”. Some attributes such as cities or states take on multiple attributes.

To find the number of frequent itemsets without having a specified minsup value, we must consider all possible itemsets. Thus, we are essentially finding all the permutations of the various attributes.

How many association rules are possible?

Describe how the support and confidence thresholds will help to prune this space.

The support signifies the *frequency* of an itemset in a database. It is defined as the number of times which the itemset appears over the total number of itemsets.

Support can help prune this space by deciding which itemsets are more common than others, pointing to various correlations between the 11 discrete attributes.

The confidence shows the likelihood of an item Y belonging to the same itemset as item X. This can be a probability P($Y|X$). It can provide some useful information as to the probability of item X corresponding with another item Y, but it does not show the frequency of item Y.

We can use confidence to prune this space by checking the conditional probabilities of the 11 discrete attributes to see which ones have a certain probability threshold over some prespecified level. We can use the $\mathrm{P}(Y|X)$ formula to find this value and compare with our threshold. If it is greater than our threshold, we can then state that the confidence of a particular item Y in our data has a probability above our threshold that item X will also exist in the same itemset as Y.

Which threshold will have a larger impact on the efficiency of the association rule algorithm?

Generally, the confidence of an item will have a larger impact on the efficiency of the association rule algorithm. Association rule learning is based on discovering interesting relations between variables in databases. It attempts to determine rules discovered in databases using some metrics of "interestingness".

The confidence threshold does exactly this – it shows the probability of an item Y existing with an item X. It does not show the *frequency* of item Y but does provide information about the association of item Y with other items in the database.

3. Implement the Apriori association rule algorithm.

See *association_rules.py*

4. Apply the Apriori association rule algorithm to the yelp5.csv data. Use cutoff thresholds of minsup = 25% and minconf = 75%.

List the 20 discovered association rules with goodForGroups=1 or goodForGroups=0 as a consequent, and largest support values, to characterize the data.

(i)

Itemset	Consequent	Support	Confidence
open	goodForGroups	0.80	0.91
waiterService	goodForGroups	0.58	0.92
caters	goodForGroups	0.45	0.92
state_AZ	goodForGroups	0.38	0.91
alcohol_full_bar	goodForGroups	0.38	0.97
alcohol_none	goodForGroups	0.39	0.84
noiseLevel_average	goodForGroups	0.63	0.92
attire_causal	goodForGroups	0.88	0.91
priceRange_1	goodForGroups	0.39	0.86
priceRange_2	goodForGroups	0.46	0.94
open_state_AZ	goodForGroups	0.33	0.92

open noiseLevel_average	goodForGroups	0.56	0.92
waiterService alcohol_full_bar	goodForGroups	0.26	0.97
waiterService priceRange2	goodForGroups	0.56	0.92
caters attire_casual	goodForGroups	0.31	0.95
noiseLevel_average state_AZ	goodForGroups	0.44	0.91
attire_casual alcohol_full_bar	goodForGroups	0.27	0.93
priceRange_2 alcohol_full_bar	goodForGroups	0.35	0.97
alcohol_none attire_casual	goodForGroups	0.28	0.97
Alcohol_none priceRange_1	goodForGroups	0.38	0.84
priceRange_1 attire_casual	goodForGroups	0.29	0.84
priceRange_1 attire_casual	goodForGroups	0.38	0.86
alcohol_none priceRange_1 attire_casual	goodForGroups	0.29	0.84

(ii)

(a) Open (no roof) places with good waiter service, caters, alcohol, and average noise level provide a good setup for groups of people to enjoy.

(b) For groups, places that are less expensive and fall in the price range of 1 and 2 are better as there is not as high of a cost to support a greater number of people.

(c) Arizona includes a lot of touristy locations, and areas of high tourists tend to have more casual restaurants. This could be because people generally do not dress up much on vacations.

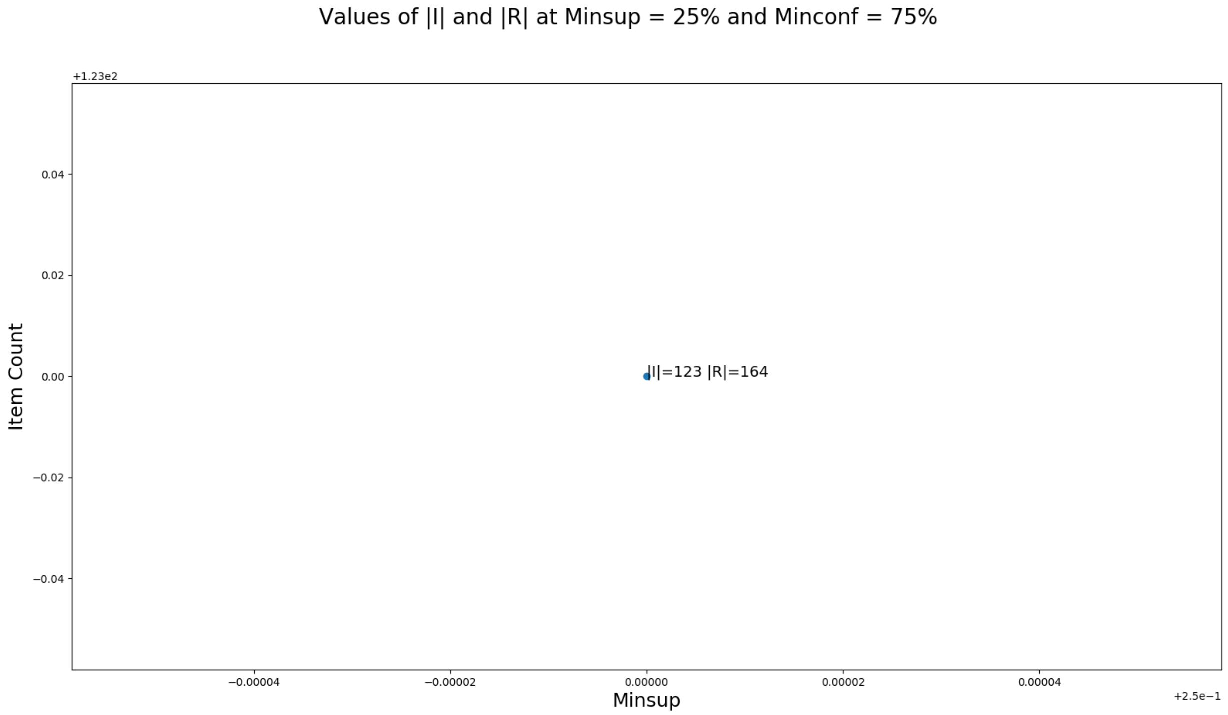
(d) A combination of the above show a much better representation of patterns and interesting correlations.

5.

Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

(i)

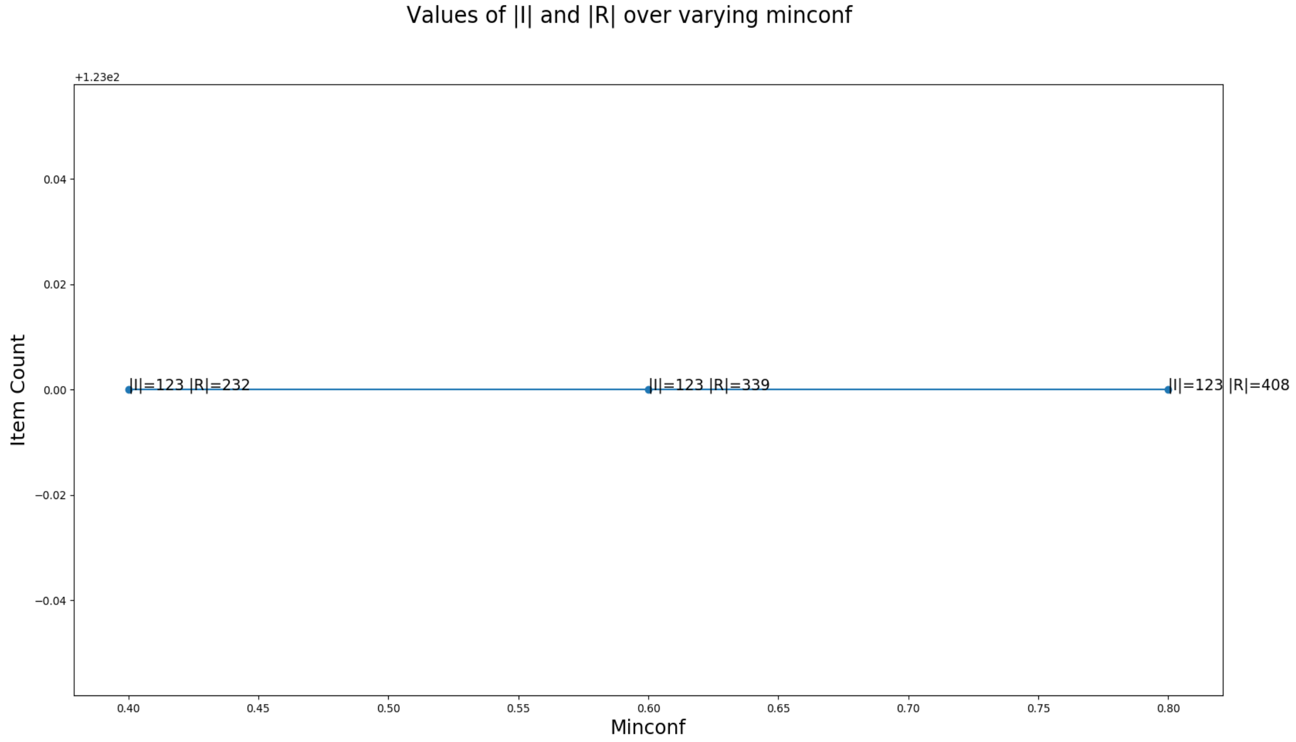
Here we keep $minsup = 25\%$ and $minconf = 75\%$



Here we can see that the number of Frequent Itemsets ($|I|$) is 123 and the number of rules ($|R|$) discovered are 164.

(ii)

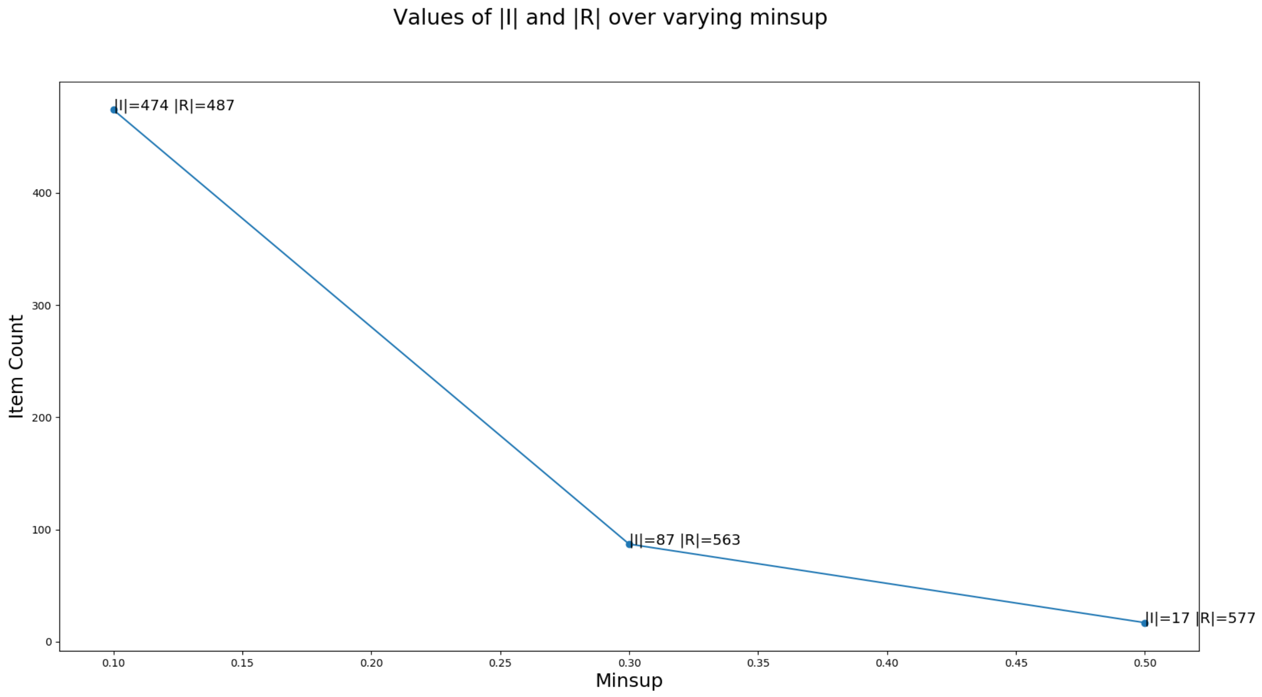
Keeping *minconf* at 75%, we can see of the values of $|I|$ and $|R|$ change as *Minsup* values change.



(iii)

Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

Keeping minsup at 25%, we can see how the values of |I| and |R| change with *minconf.*



Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019

Akash Lankala
PUID: 0027710383
CS 373
Homework 3
October 30th, 2019