Assignment 3

Date of Completion :- 17.9.2020

Date of Submission :- 29.9.2020

Title :- Apply a-priori algorithm to find frequently occuring items from given data.

Problem Statement :- Apply a-priori algorithm to find frequently occuring ~~techniques~~ items from given data and generate strong association rules using support and confidence thresholds. Eg :- Market Basket Analysis.

Learning Objectives :- To understand a-priori algorithm and its applications.

Learning Outcomes :- Students will be able to understand a-priori algorithm and predict association rules.

Software/Hardware Requirements :- Jupyter Notebook, python.

Theory :-

Association Rules :-

1) These rules help discover relationships between seemingly independent relational datasets or other data repositories.

1) Support Count:- Frequency of occurrence of a itemset.
2) Association Rule - An implication expression of the form $X \rightarrow Y$ where X and Y are any 2 itemset.
3) Frequent Itemset:- An itemset whose value is greater than or equal to minsup threshold.

Rule Evaluation Metrics:-

1) Support:- $(X + Y) \div$ total
interpreted as fraction of transaction that contain both X and Y.

2) Confidence -
$$Supp(X \cup Y) \div Supp(X)$$
Measures how often each item in Y appears in transactions that contain items in X also.

3) Lift (I)
$$(X \Rightarrow Y) = Conf(X \Rightarrow Y) \div Supp(Y)$$
Lift value near 1 indicates X & Y almost often appear together as expected, greater than 1 means they appear together more than expected & less than 1 means they appear less than expected.
Greater lift indicates stronger association

Apriori Algorithm:-

1. Used for finding frequent itemset in a dataset for boolean ~~expr~~ association rule.
2. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties.
3. An iterative approach or level wise search where

K - frequent itemsets are used to find K+1 itemset.

Limitation of Apriori Algorithm:-

1) Slow
2) space required to hold large number of candidate sets.

Dataset Used:- Market Basket

Test case / Algorithm / Example:-

TID	items
T1	I_1, I_2, I_5
T2	I_2, I_4
T3	I_2, I_3
T4	I_1, I_2, I_4
T5	I_1, I_3
T6	I_2, I_3
T7	I_1, I_3
T8	I_1, I_2, I_3, I_5
T9	I_1, I_2, I_3

Support = 32% Confidence 60%

I

item	freq	Support %.
I_1	6	6/9 = 0.67 => 67%
I_2	7	7/9 = 0.78 => 78%
I_3	6	6/9 = 0.67 => 67%
I_4	2	2/9 = 0.22 => 22%
I_5	2	2/9 = 0.22 => 22%

Select items on support > 32% I_1, I_2, I_3

II

item	freq	support
I_1, I_2	4	$4/9 = 0.44 \Rightarrow 44\%$
I_1, I_3	4	$4/9 = 0.44 \Rightarrow 44\%$
I_2, I_3	4	$4/9 = 0.44 \Rightarrow 44\%$

Rule	Support	Confidence	Confidence %
$I_1 \rightarrow I_2$	4	$4/6 = 0.67$	67
$I_2 \rightarrow I_3$	4	$4/7 = 0.57$	57
$I_1 \rightarrow I_3$	4	$4/6 = 0.67$	67
$I_3 \rightarrow I_1$	4	$4/6 = 0.67$	67
$I_2 \rightarrow I_3$	4	$4/7 = 0.57$	57
$I_3 \rightarrow I_2$	4	$4/6 = 0.67$	67

Association Rules accepted

$I_1 \rightarrow I_2$

$I_1 \rightarrow I_3$

$I_3 \rightarrow I_1$

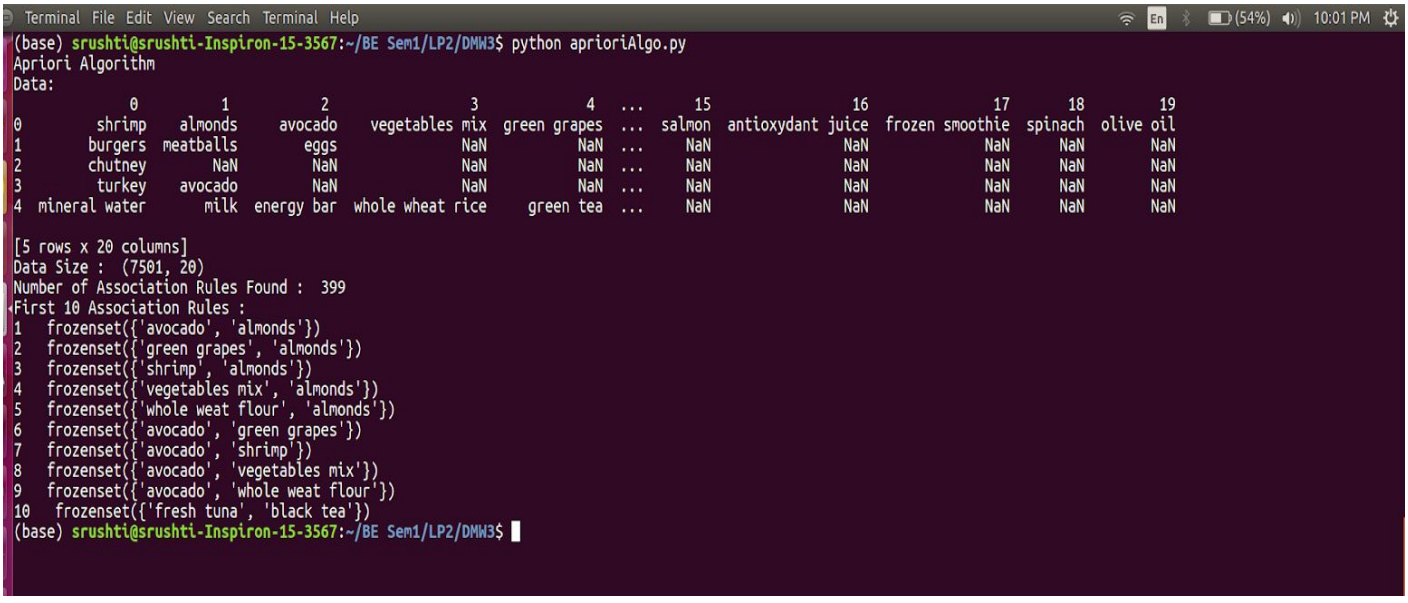
$I_3 \rightarrow I_2$

Conclusion: Thus we have understood and implement apriori algorithm to find the association rule in dataset.

CODE:

```
import numpy as np
from apyori import apriori
import pandas as pd

print("Apriori Algorithm")

data=pd.read_csv('Market_Basket_Optimisation.csv', header = None)

print("Data:")
print(data[:5])
print("Data Size : ",data.shape)

records=[]

for i in range(0,22):
        records.append([str(data.values[i,j]) for j in range(0,6)])

association_rules=apriori(records, min_support = 0.005, min_confidence = 0.5, min_lift = 3, min_length = 2)
association_result=list(association_rules)

print("Number of Association Rules Found : ",len(association_result))
print("First 10 Association Rules : ")
for i in range(0,10):
        print((i+1)," ",association_result[i].items)
```

OUTPUT:



Apriori Algorithm

Data:

	0	1	2	3	4	...	15	16	17	18	19
0	shrimp	almonds	avocado	vegetables mix	green grapes	...	salmon	antioxydant juice	frozen smoothie	spinach	olive oil
1	burgers	meatballs	eggs	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
2	chutney	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN
3	turkey	avocado	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN

[4 rows x 20 columns]

Data Size : (7501, 20)

Number of Association Rules Found : 399

First 10 Association Rules :

1 frozenset({'avocado', 'almonds'})
2 frozenset({'green grapes', 'almonds'})
3 frozenset({'shrimp', 'almonds'})
4 frozenset({'vegetables mix', 'almonds'})
5 frozenset({'whole weat flour', 'almonds'})
6 frozenset({'avocado', 'green grapes'})
7 frozenset({'avocado', 'shrimp'})
8 frozenset({'avocado', 'vegetables mix'})
9 frozenset({'avocado', 'whole weat flour'})
10 frozenset({'fresh tuna', 'black tea'})