Data Science and Analytics (IT3080)

B.Sc (Hons) in IT (IT Specialization)

Year 3

SLIIT
Discover Your Future

Lab Sheet 02

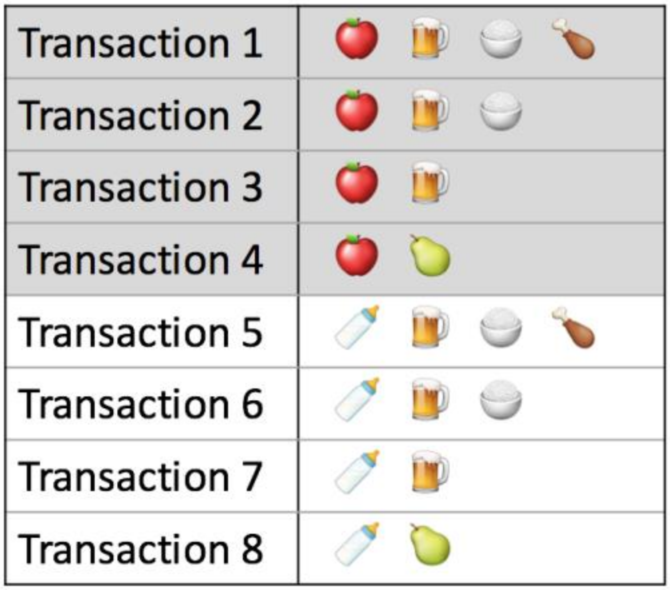
Pre-requirements

Launch Anaconda Prompt from the Start Menu on Windows. Type “pip install apyori” on the command prompt to install required libraries for using the Apriori Algorithm.

Key-terms to remember

Support: Support refers to a way to express how often item sets are occurring in baskets.

Ex : For example consider the following transactions



Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

$$\text{Support}\ \{🍎\} = \frac{4}{8}$$

Confidence: Confidence is used to determine whether one or more items in the set are directing the purchase of other items.

For ex:

$$\text{Confidence}\ \{🍎 \rightarrow 🍺\} = \frac{\text{Support}\ \{🍎, 🍺\}}{\text{Support}\ \{🍎\}}$$

Activity A

Download Apriori.ipynb and store_data.csv from courseweb.

Create a folder under the notebook and upload the downloaded files to the notebook.

Open Apriori.ipynb and go through the code line by line. Following section explains the coding.

```
import pandas as pd

from apyori import apriori
```

The pandas library is an important library in python used for data analysis. It contains tools for reading and writing data flexible, objects to store data for processing & etc. In this session we are going to use the apriori function avaialable in apyori module in python.

```
store_data = pd.read_csv("store_data.csv", header=None)

display(store_data.head())
```

The lines above are importing the dataset from the csv file and converting it into a pandas data frame so that we can access the data easily. Note that we don't have a header labeling our data so that we can set it in parameters. Head method is used to retrieve first 5 rows of the data frame.

```
records = []

for i in range(1, 7501):

  records.append([str(store_data.values[i, j]) for j in range(0, 20)])
```

The next step is very important, apriori algorithm takes the input as list of lists, so we need to make our dataset into a list of list format, the nested loop will do the job for us. After these lines the transaction will have list of transactions, each transaction will have the items that were purchased.

```
association_rules = apriori(records, min_support=0.0045, min_confidence=0.2,

                                                  min_lift=3, min_length=2)

association_results = list(association_rules)
```

The next step is to actually call the apriori function with parameters of minimum support, minimum confidence, minimum lift and the minimum length of the itemset required. The function will return a result set. This result is converted to a list first and then the content of the list are printed.

Execute the coding and observe the outputs.

Change the min_support, min_confidence and min_lift parameters for the apriori function and observe the changes in the output received.

Activity B

Download the “movie_dataset.csv” from courseweb. Open the file and observe the contents. The file contains information of movies purchased from an online movie store. Each row in the file shows a single transaction.

Develop a python code which uses Apriori algorithm to do the following

1. Find frequent itemsets with two items having minimum support of 1%.
2. Find all association rules with two items having minimum support of 1% and minimum confidence of 50%.