CS F415: DATA MINING

ASSIGNMENT 1: REPORT

Goal: Generate frequent itemsets and interesting association rules using apriori algorithm

Name : NEETU
ID: 2015A7PS0079H

Dataset used: Groceries Market Basket Data (groceries.csv)
www.sci.csueastbay.edu/~esuess/classes/Statistics_6620/Presentations/ml13/groceries.csv

Programming language : Python

Files submitted:

*	src code file : apriory.py
*	association rules generated: rules.txt (sup = 30, conf = 0.01)
*	frequent itemsets generated: frequent_itemsets.txt (sup = 30, conf = 0.01)
*	Report
*	Readme.md

Preprocessing done on the data:

Dataset (groceries.csv) had data present in transactions format (items present in each transaction were seperated by commas).

Preprocessing included:

*	stripping the data of seperators (cleaning)
*	identifying transactions and storing them seperately

- getting rid of duplicate data

Formulas used:

Support

The support of an itemset is the proportion of transaction in the database in which the item X appears. It signifies the popularity of an itemset.

$$supp(X) = \frac{\text{Number of transaction in which} X \text{appears}}{\text{Total number of transactions}}.$$

Confidence:

It signifies the likelihood of item Y being purchased when item X is purchased

$$conf(X \longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Number of frequent itemsets and association rules for different values of support and confidence :

1. Support = 45, confidence = 0.03
Number of rules = 2476
Number of frequent itemsets = 1192

2. Support = 30, confidence = 0.01
Number of rules = 5232
Number of frequent itemsets = 2226

3. Support = 55, confidence = 0.05
Number of rules = 1561
Number of frequent itemsets = 857