CS 6103D Software Systems Laboratory
Assignment 2a: Learning Python Basics

October 10, 2018

Learning Objective

- Learn operations on *Strings*, *Lists*, *Dictionaries*
- Learn to use *re* module for regular expressions
- Functions and Functional programming, use of *lambda* Operator

Problem Description

The Market Basket file uploaded (Market_Basket_2.csv) is a modified version of the Kaggle dataset https://www.kaggle.com/c/instacart-market-basket-analysis/data. It has a set of transactions made at a supermarket. Download and use the dataset to program the following:

1. Read the file into a *List* of *Dictionary* with 'TID' as the *Key* and 'CartItems' as the *Value*. You can store 'CartItems' as a String.
2. Add a step in the program to convert the 'CartItems' into a *List*
3. Use 're' module to get all specific 'Items' like 'Whole Bread', 'Dessert Wine', 'Whole Milk' etc. and convert the relevant ones to more generic 'Items' like 'Bread', 'Wine', 'Milk' etc. Use a *lambda* function to make the transformation. Once you assess the data for such occurrences, you can maintain a map to do the transformations.
4. Store the modified 'CartItems' in lexical order
5. Generate a list of frequent items - items which appear in at least 10% (support factor) transactions
6. Combine the frequent items with length one to create a 2-itemSet (a list with two frequent items) with a minimum support factor of 10% . While generating k-itemsets from (k-1)-itemsets, merge them only if the first k-2 items are the same. Append the (k-1)-th item from both the lists to the k-2 common items to generate k-itemsets.
7. Iterate the above procedure to generate all frequent itemsets with a minimum support factor of 10%. Store the generated frequent n-itemSets into a file.