資料探勘 Homework1

學生：吳汶峻

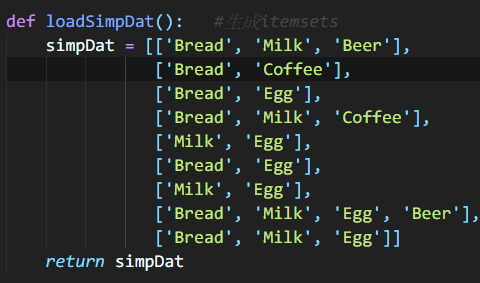
學號：Q36071229

系級：電通甲

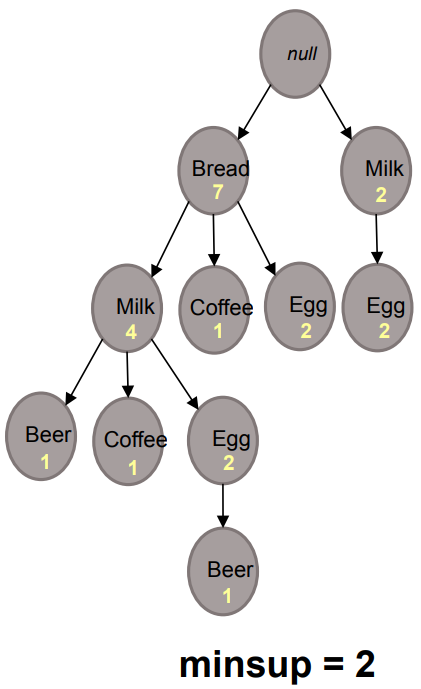
一、用Sample Data跑程式

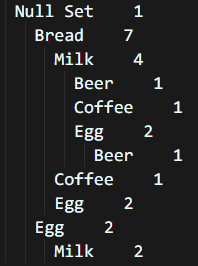
首先，在我github上FP\_Growth.py的code中，我設計與老師範例(FP Growth Example.pdf)相同的Item Table內容，如下：

Sample data為：



將Min-support設定為2、產生FP-tree後，可以得知範例的Tree-node與count跟code執行結果相同：

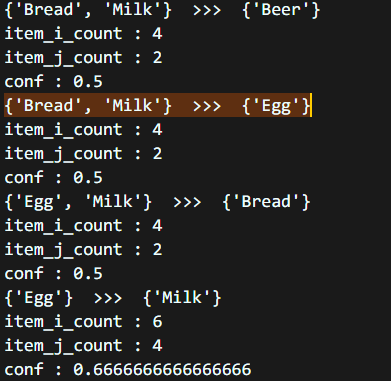




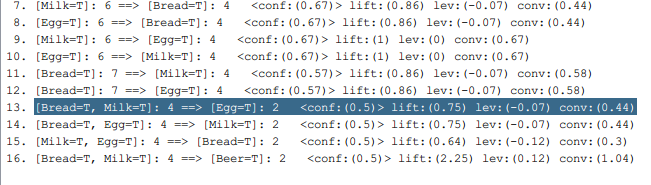
(此顆Tree是由左到右生成，最左邊(Null Set)為Root，左右距離相同的位於同一層，依序往右長出。)

可以利用FP-Growth.py中的associate()函數將每個Frequent item sets與Confidence列舉出來，Frequent item sets與其Confidence執行結果如下：

(Min-support=2的Frequent item sets)



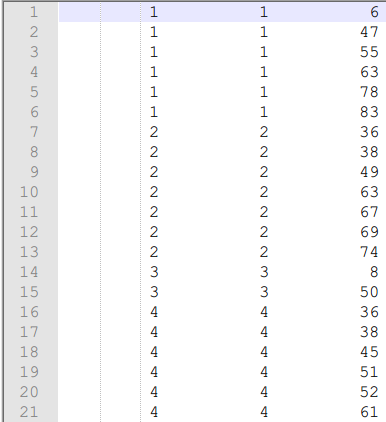
接著再將Data放進Weka比較其關聯法則(Associate-Rule)及Confidence：

在圖中可以發現反白處的關聯法則與Confidence是一模一樣的，其他關聯法則及Confidence在做比較後發現也是一模一樣！

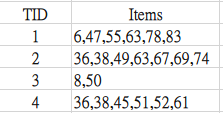
二、IBM Quest Synthetic Data Generator

接著，利用老師提供的IBM Quest Synthetic Data Generator產生data數據，這個產生器可以依據自己喜好來選擇想要的data長度、數量及平均每筆item數量，而在我的實作中，我使用以下指令：

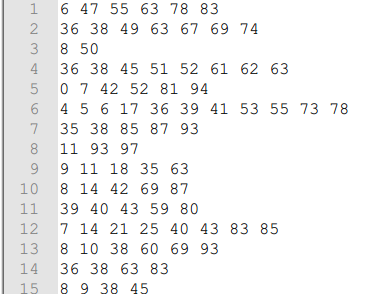
「"IBM-Quest-Data-Generator.exe" lit -ntrans 0.1 -tlen 5 nitems 0.1」

目的是要產生ID總數85、平均每個ID有5個items，item範圍是0~100的data。跑出來的部分data如下：

看了很久，再來比較老師PPT的內容，可以推算出可能為下列情形：

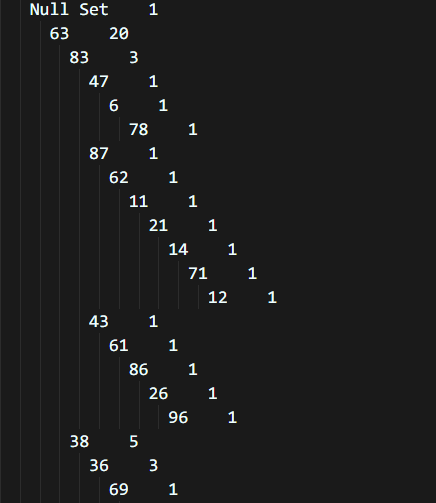
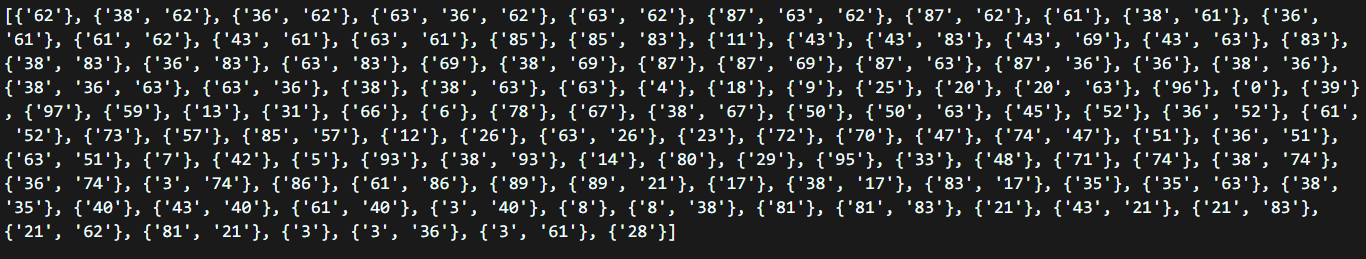


但這個data格式不是我想要的，所以寫了一個Data\_conversion.py的檔案來讓資料變得更簡潔：

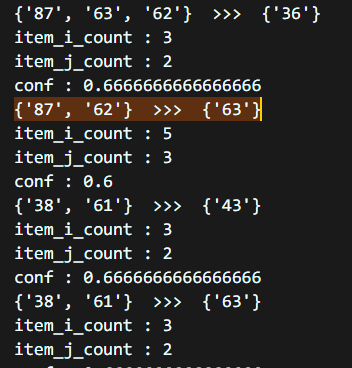


在上面data中，左邊深色背景為TID，右邊為其Items，這樣就可以進行FP-Growth的動作了。

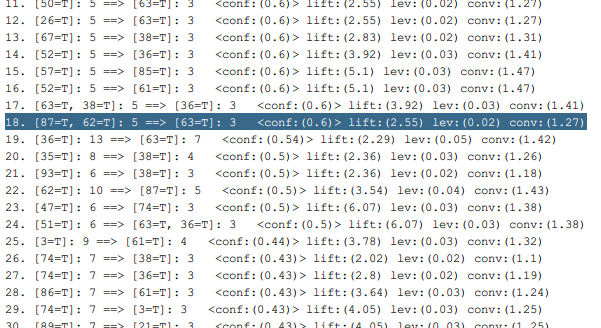
三、執行FP-Growth程式與比較結果

接著，我用Data\_read.py開始讀檔並利用FP-Growth.py進行資料剖析，我將min-support設定為3，找出FP-tree還有其Frequent Item sets：

(上圖為min-support設為3的Frequent Item sets)

然而，由於我使用的是Weka這個資料分析工具，將Data丟進Weka跑時產生的是關聯法則(Associate-Rule)及其Confidence，所以我再利用先前產生的Frequent Item sets產生出關聯法則及每個Confidence，其部分結果為：

而Weka的設定值min-support=3.0、Confidence=0.5，跑出來結果為：



在比較Code與Weka出來的結果，可以發現在圖中反白的畫面其Confidence都是0.6，而其他的數據也可以從程式執行結果找出。由此可知，程式跑出來與Weka執行結果是一模一樣的！