資料探勘 Homework1

學生：吳汶峻

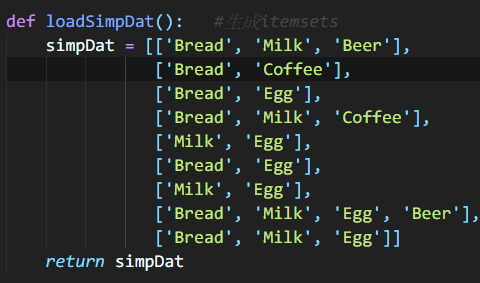
學號：Q36071229

系級：電通甲

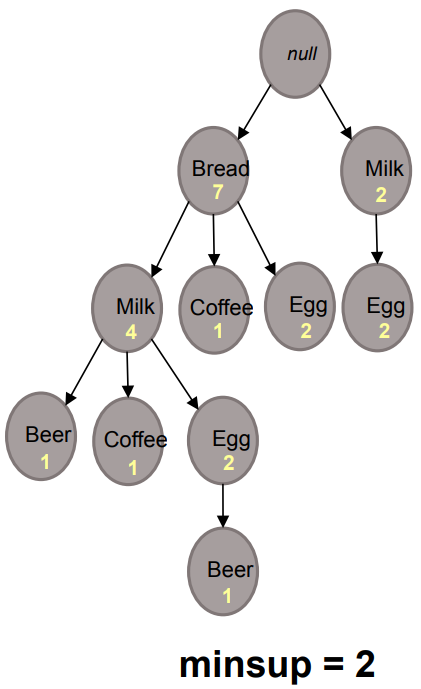
程式環境：Python3/Windows10

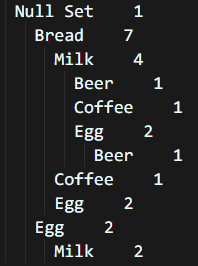
一、用Sample Data跑程式

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跟程式執行結果相同：

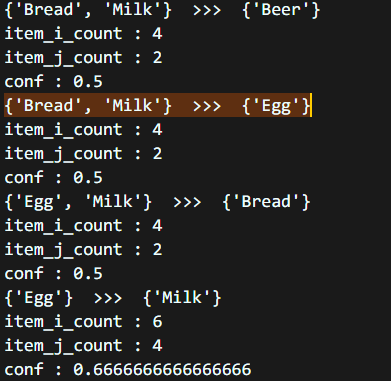




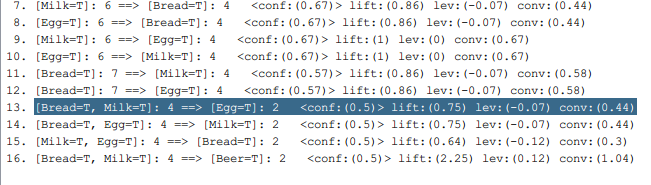
(此顆Tree是由左到右生成，最左邊(Null Set)為Root，左右距離相同的位於同一層，依序往右長出。)

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)



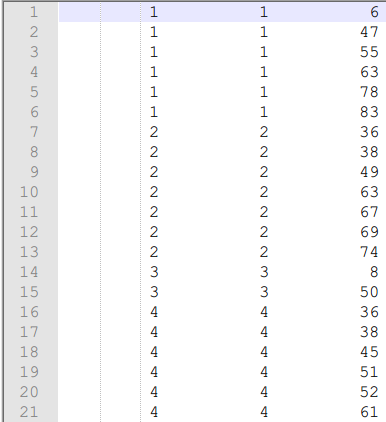
接著再將Data放進Weka比較其關聯法則(Associate-Rule)及Confidence：

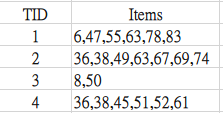
在圖中可以發現反白處的關聯法則與Confidence是一模一樣的，其他關聯法則及Confidence在做比較後發現也是一模一樣！

二、IBM Quest Synthetic Data Generator

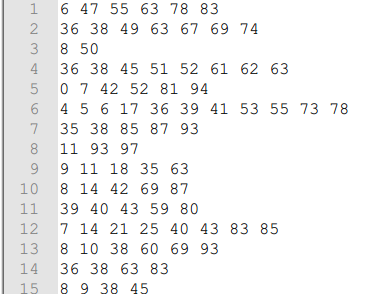
接著，利用老師提供的IBM Quest Synthetic Data Generator產生data數據，這個產生器可以依據自己喜好來選擇想要的data長度、數量及平均每筆item數量，而在我的實作中，我使用以下指令：

「"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：

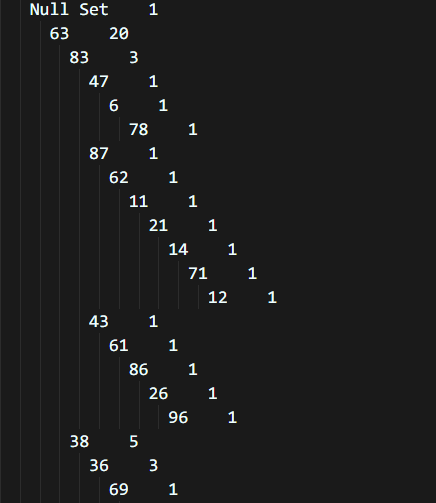
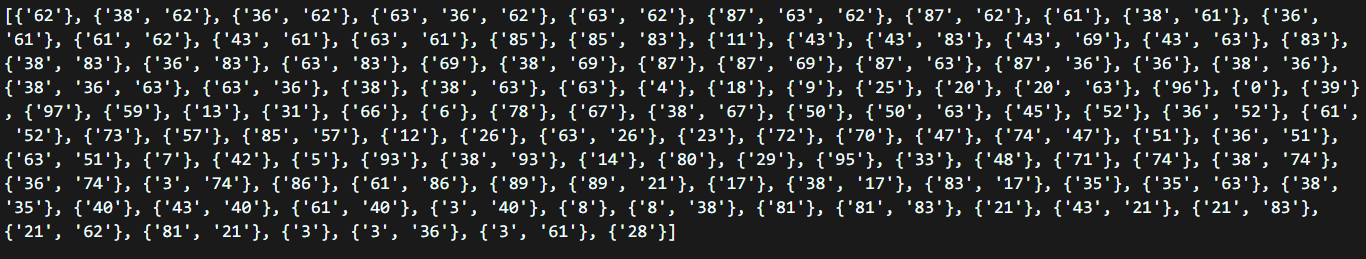
在思考上述Data想表達的涵義，並比較老師PPT的內容，可以推導出可能為下列情形：

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：

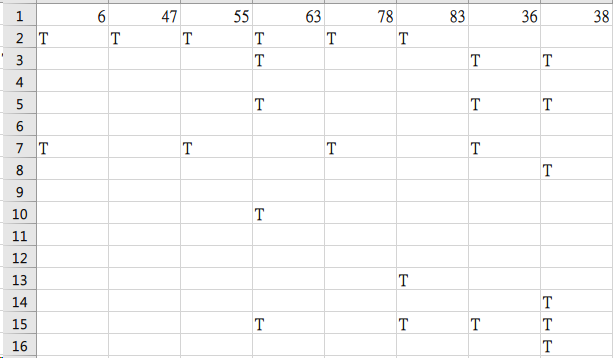


在上面data中，左邊深色背景為TID，右邊為其Items，這樣就可以進行FP-Growth的動作了。

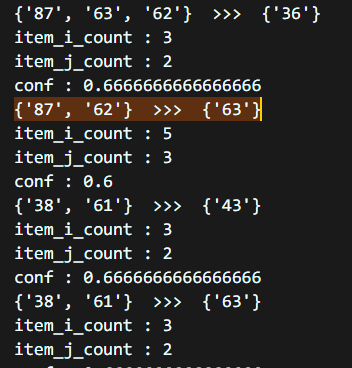
三、執行FP-Growth程式與比較結果

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：

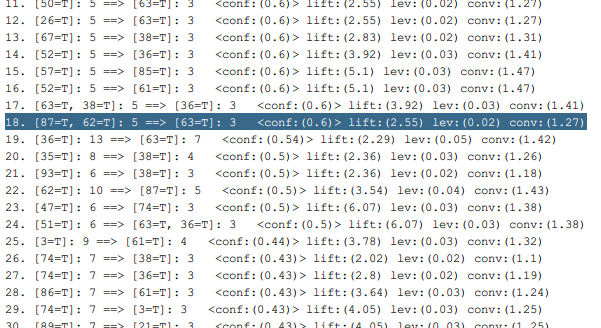
(上圖為min-support設為3的Frequent Item sets)

然而，由於我使用的是Weka這個資料分析工具，它所需要的CSV檔格式中第一列必須為”屬性名稱”，因此我將Data利用Data\_convert\_Weka.py的程式轉成Weka承認的格式：

其中空白的部分為該TID沒有此數值(Null)，T代表此TID有此數值。

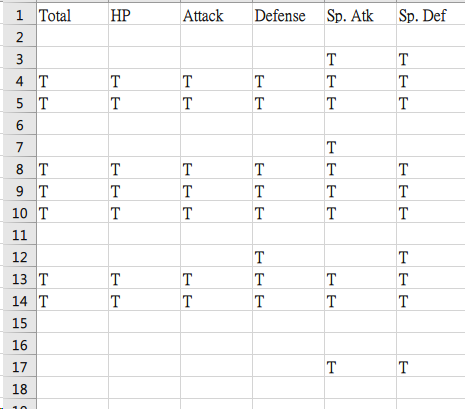
而在Data丟進Weka跑時產生的是關聯法則及其Confidence，所以我再利用先前產生的Frequent Item sets生成關聯法則及每個Confidence，其部分執行結果為：

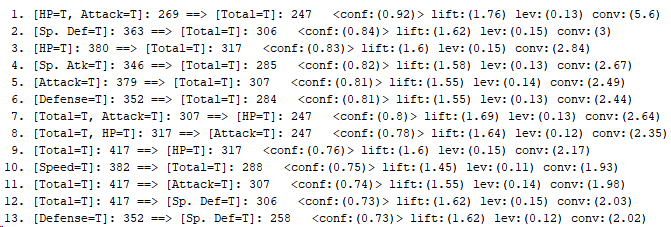
而將Weka的設定值設為min-support=3.0、Confidence=0.5，跑出來結果為：



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！

四、Kaggle

 在www.kaggle.com的網站裡找到一個適合做Weka的Data(Pokemon with stats)，這是一個pokemon各種屬性值(Ex:HP、Attack…)的Data。下載下來後，經過Data Parsing，將極端值及獨立的數據刪除後，有分析出與Weka符合的格式：

Weka設定將min-support=2、confidence>0.5，執行結果如下：

有顯示出關聯法則及其Confidence，但因為這份Data若要放在我的程式執行的話，需要分析更多的Data parsing才能夠建FP-tree，目前Kaggle Data分析只做到Weka實際分析的部分。

五、實作心得

剛開始聽到老師說要做FP-Growth的實作時，當下是很緊張的，擔心自己做不出來，但在自己查找資料、了解FP-Tree涵義是什麼後，其實只要肯花心思與時間去學習跟寫code，就能學到很多有關寫程式技巧與演算法。

而在IBM程式產生Data、跑入我的程式與Weka所展示的結果相符合後，我在Kaggle上雖然有找到一個關於pokemon特性的比較，蠻適合當Weka分析的Data，但由於它的Data parsing有點難去分析，至今尚未能在我的程式中順利執行，還沒想到較好的解決方法，不過我會在往後的時間盡可能將其分析出來，將Kaggle的Data做出FP-Growth，因為這是實作有魅力又能學到很多技巧的地方！