***Data Mining Project 1 Report***

**資訊所碩一 P76074575 潘昌義**

1. **作業簡述**
2. **資料集的蒐集、製作與評估**
3. **Apriori演算法**
4. **FP-growth演算法**
5. **總結**
6. **作業簡述**

在尋找關聯法則的課程中，我們學到兩種雖然古早但不失經典的兩種演算法：Apriori演算法與FP-growth演算法。其中Apriori演算法為一項集大小尋找每一個項集的出現次數後，與minimum support進行比較好決定是否要刪除較不重要（或者較弱）的關聯法則；FP-growth演算法則是先找一次單一項的數量與minimum support比較後，直接建一棵FP-tree將其每個單一項的資料存入樹裡，並以link將樹中相同資料的node串接起來，方便未來追蹤該值曾經在什麼地方出現過。本次作業的目的是希望我們實際撰寫一遍Apriori演算法與FP-growth演算法，並透過不同資料集的測試，來比較這兩個演算法在各方面的差異，例如速度或者對資料複雜的容忍度等。

1. **資料集的蒐集、製作與評估**
2. IBM Quest Data Generator

這個資料集是上課指定一定要用的模擬資料，其生成資料大小可透過改變參數來決定，以下列出我最後有修改的參數：

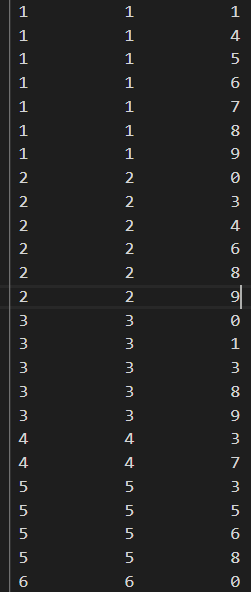
-ntrans: 決定交易筆數，標準為1000

-tlen: 決定每筆交易內大概有多少個物件，會上下起伏+2~-2

-nitems: 決定一共有多少種不同的物件會出現在交易清單裡

為了讓資料呈現”100筆交易，每筆交易約5個物件，物件種類10個”，我最後調整如下，並附上前25筆資料生成的結果：

lit -ntrans 0.01 -tlen 5 -nitems 0.01



1. Dataset from Kaggle
2. **Apriori演算法**