Report

Dataset

使用的Dataset來自於IBM Quest Data Generator，由於可以自定義transaction的個數以及item的數量，所以此次作業使用一筆較小的資料下去跑，另一方面也是方便debug的緣故所以做此選擇。

程式

這次要求實作用FP-growth找出association rules，還要比較測試資料的結果，就理論來說FP-growth完成找出association rules的速度會比較快，因為只需要掃data兩次即可，但因為測資較小所以可能看不出差別，不過肯定是比暴力法每次掃data要快上許多，詳細的步驟如下:首先我是使用python，當把dataset建立完之後，首先建立一顆FP-tree，主要分成兩部分，第一部分掃描database，統計出所有的frequent item。第二部分再掃一次，把小於minimum support的元素替除掉，然後把frequent item做sorting，把item插入fp-tree之中，不斷地更新fr-tree。接下來要開始mining frequent item，對每個元素找到他的prefix path，之後將所有prefix path做為新的data set 傳入創造fp-tree這個function，稱做一個conditional fr-tree，上述兩步持續重複到fp-tree之中只剩一個元素。程式的概念大約如上。

Association Rules

Spec中有提到使用weka來進行關聯法則的尋找，遺憾的是嘗試了許多方法，仍無法讓weka使用generator產生出來的data進行分析，似乎是因為generator產生出的隨機data全是數字導致，雖然沒有完成，但仍需在此報告一下。