

國立成功大學 資工所

Data Mining

資料探勘

**Project 1**

**Association Analysis**

課程教授: 高宏宇

學生: 葉芯妤

學號: P96074147





目錄

一、專案內容簡介 ……………………........…………3

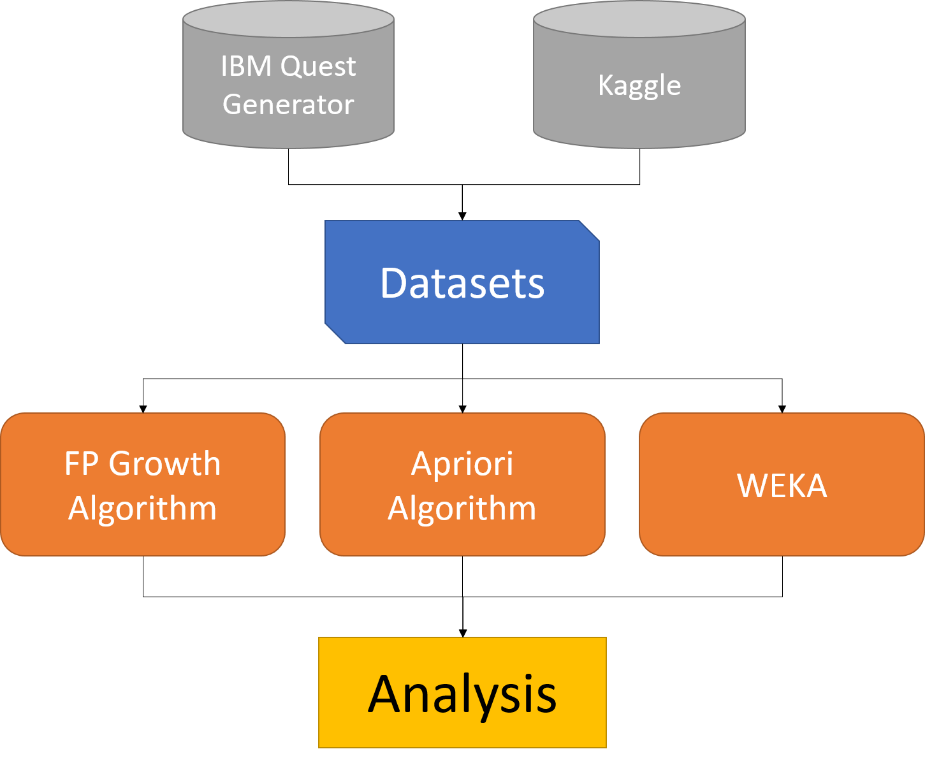
二、程式實作 ……………………................…………4

三、WEKA分析 …………...........................…………6

四、結論 …………………........................……..……..9

1. 簡介
2. **總覽**

利用IBM資料產生器產生的資料集以及從Kaggle網站上找到可用於關聯分析的資料集，實作Apriori演算法以及FP-Growth演算法的程式碼並使用WEKA資料探勘工具來分析兩項資料集的結果，專案架構如圖一所示。



圖一：專案架構

1. **資料**

本專案利用IBM資料產生器設定不同變因的參數作為輸入，並產生多種測試資料集，對於不同資料集能分析的東西也相對較多。更改IBM-Quest-Data-Generator.exe參數產生資料，使用lit -ntrans 10 -tlen 50 -nitems 0.1指令，得到dataset後，將檔案轉為csv檔，以利分析，資料內容如下圖二所示。

而在Kaggle網站上找尋與關聯分析相關的資料集，本專案使用國際貿易公約（CITES）野生動植物交易資料庫，分析野生動植物交易的類型以及用途之間的關聯。

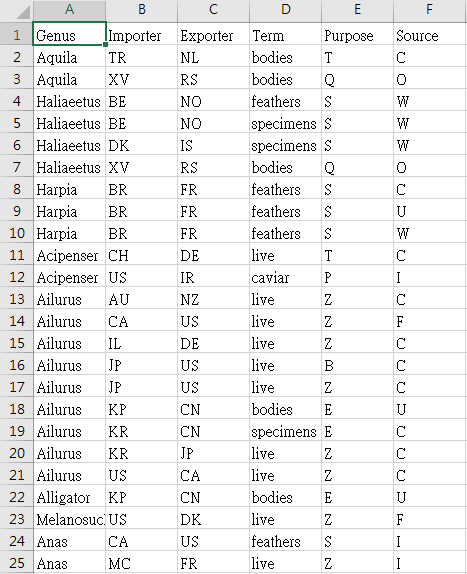
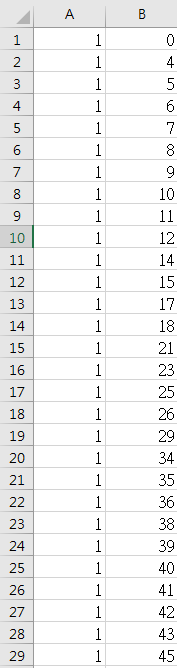
Kaggle資料來源:

<https://www.kaggle.com/cites/cites-wildlife-trade-database>

此資料為2016年和2017年進行的CITES保護物種的所有合法國際貿易，資料總數共671621筆，本專案從中隨機取樣5000筆資料進行分析，資料內容如下圖三所示，而表格一為每一欄資料所代表的意思，其中有兩項為字母代碼，其涵義需參閱資料所附的CITES文檔。

CITES文檔網址：

<https://trade.cites.org/cites_trade_guidelines/en-CITES_Trade_Database_Guide.pdf>



圖二 IBM產生的資料內容 圖三 Kaggle資料內容

表一 Kaggle資料欄位說明

|  |  |  |
| --- | --- | --- |
| Genus | 學名 | 代表物種的分類 |
| Importer | 進口商 | 進口該物種的國家 |
| Exporter | 出口商 | 出口該物種的國家 |
| Term | 型態 | 出口或進口的型態。活體、標本、獸皮等。 |
| Purpose | 目的 | 物種出口或進口的目的。為字母代碼，其涵義需參閱CITES文檔。 |
| Source | 來源 | 描述動植物如何進入市場：繁殖，捕獲野生動物，標本等。為字母代碼，其涵義需參閱CITES文檔。 |

1. 程式實作
2. **環境**

作業系統：Window 10

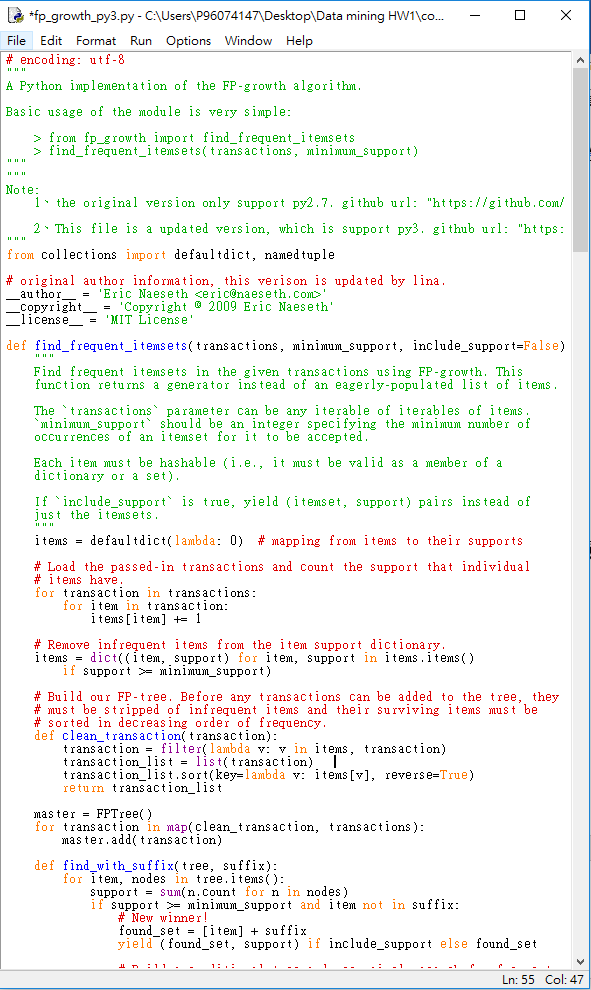
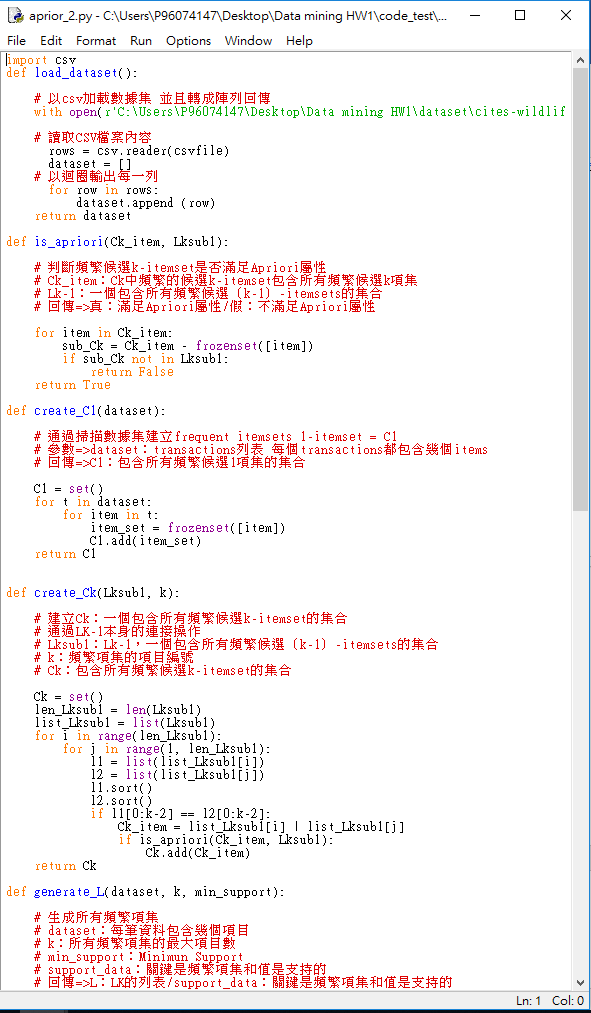
程式語言：Python3.6

額外工具：WEKA、IBM-Quest-Data-Generator、Excel、Kaggle

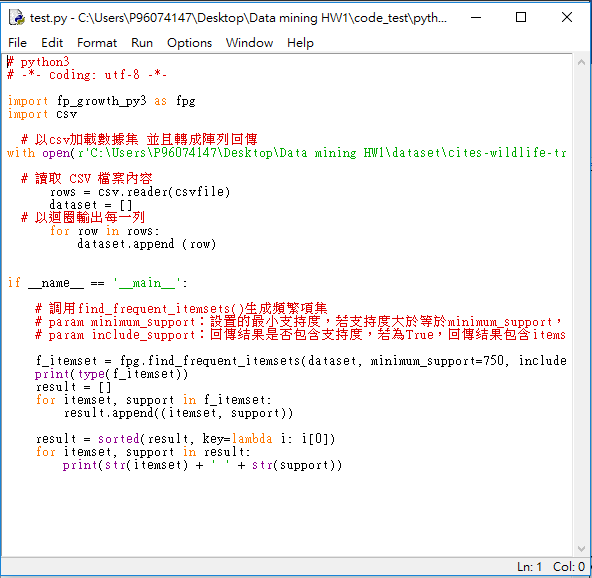
1. **程式碼**

本專案以Python撰寫Apriori演算法以及FP-Growth演算法的程式碼，對IBM及Kaggle兩項資料進行分析。由於個人程式能力不足，因此FPgrowth的程式碼是在<https://github.com/enaeseth/python-fp-growth>的基礎上做更改。

下圖四為Apriori程式碼，紅色中文字為對程式碼所作的註解。而FPgrowth程式碼分為兩部分，一部分為FPgrowth的規則，即為上述提及參考的程式碼，如下圖五所示，另一部分為個人撰寫的應用該規則，將csv檔案匯入進行分析，並顯示出結果的程式碼，如下圖六所示，紅色中文字為對程式碼所作的註解。

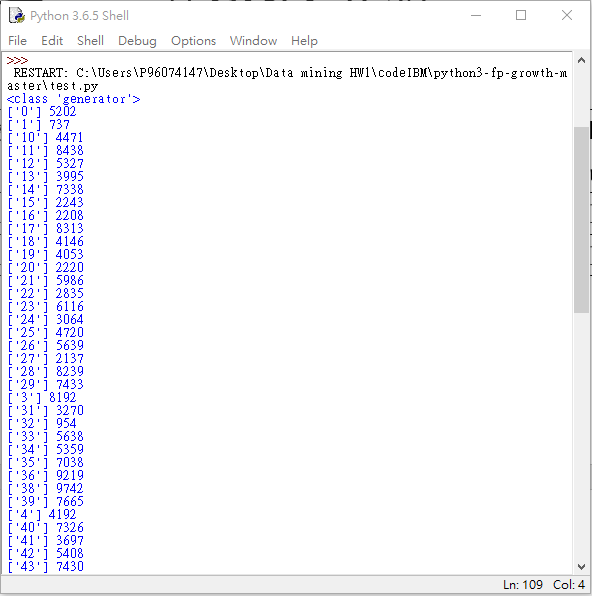


　　　　　　圖四為Apriori程式碼　　　　圖五FPgrowth參考程式碼

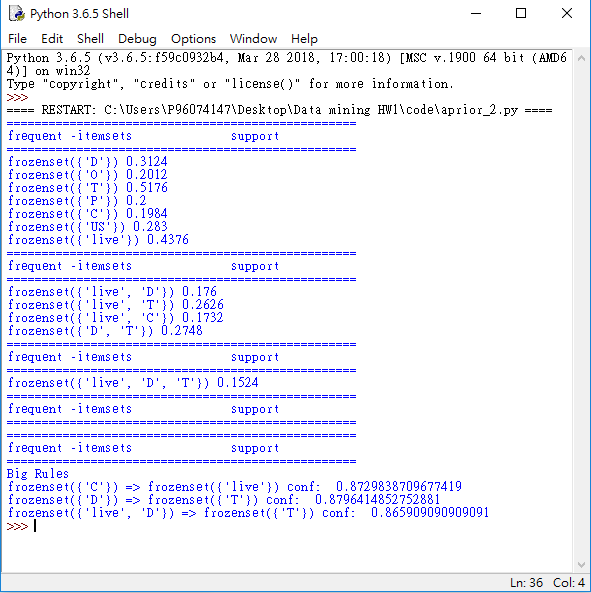
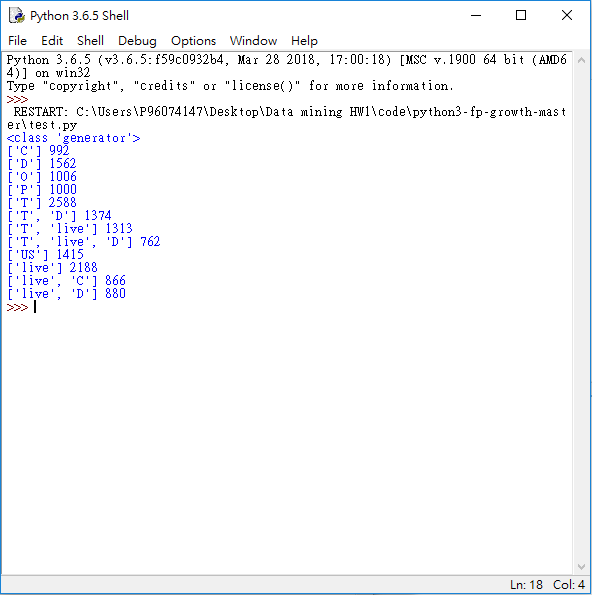


圖六　FPgrowth實際測試程式碼

由於IBM所產生資料數量相當大，在利用Apriori演算法運算時，電腦速度變得相當緩慢，且無法運算出規則，耗費相當多時間，由於FPgrowth演算法運算比Apriori快速許多，故採用FPgrowth演算法對IBM資料進行分析比較，程式執行結果如下圖七所示。



圖七　FPgrowth程式對IBM執行結果

****

圖八　Apriori程式對Kaggle執行結果 圖九　FPgrowth程式對Kaggle執行結果

1. WEKA分析
2. IBM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 交易筆數 | 平均每筆交易物品量 | 物品種類數量 | Minsup | FP Growth找到的規則量 | 執行時間 |
| 1 | 50 | 0.1 | 500 | 1426 | 69457 |
| 1 | 50 | 0.1 | 600 | 1687 | 2943 |
| 1 | 50 | 0.1 | 700 | 2106 | 944 |
| 1 | 50 | 0.1 | 800 | 1558 | 507 |
| 1 | 50 | 0.1 | 900 | 230 | 313 |

一開始將Minsup設定為500測試找尋的規則數量多寡與Minsup值的關係，由於結果冗長且每個物品項目數量出現的次數幾乎都超過500，於是將Minsup調高以100為遞增找尋各個關係。

出乎意料的，在門檻值調動後，一開始找尋到的規則數反而是增加的，而在門檻值為700時，達到2000以上後，又逐次降低。可能原因是由於門檻值調動，使物品相關的規則數找到更多。如下圖所示。

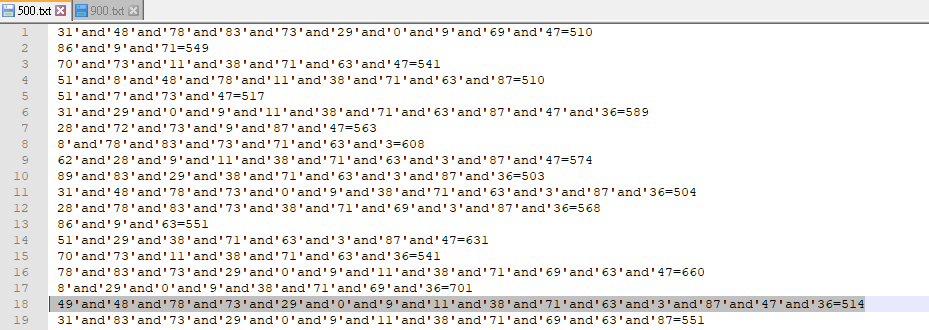


圖 Minsup為500時的結果

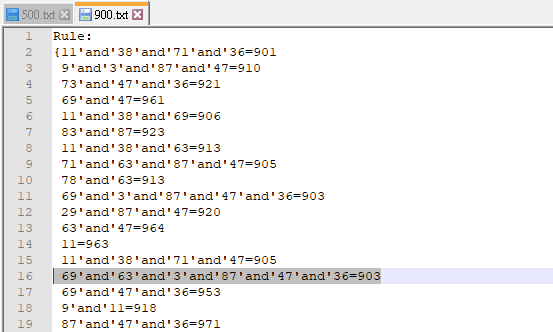


圖 Minsup為900時的結果

1. Minsup值低時，會有較長的子集

當Minsup為500時最長的一列(也就是包含最多物品數量的子集)，表示內容的物品項目任意數量皆會超過後面的數值，以18行為例，49與48常出現的比例會高過514。所以規則數量有可能會因為Minsup值的調整而有所變動。

1. 門檻值的訂定

可以看到當Minsup值定為900時，仍有許多物品集規則是超過900的，所以初訂定為500時執行時間為60000多秒，實際上是因為Minsup值極低，導致運算量大，且規則非常詳細。因此，本專案調整了物品數量以及Minsup值，更改後的資料如下表所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 交易筆數 | 平均每筆交易物品量 | 物品種類數量 | Minsup | FP Growth找到的規則量 | 執行時間 |
| 1 | 10 | 0.1 | 100 | 125 | 203 |
| 1 | 30 | 0.1 | 100 | 2734 | 6010 |
| 1 | 30 | 0.1 | 300 | 782 | 952 |
| 1 | 10 | 0.1 | 100 | 125 | 203 |
| 1 | 30 | 0.1 | 100 | 2734 | 6010 |

從平均每筆交易物品量與Minsup可以看到有相當大的差異，在測試程式時，會發現運算速度與找到的規則量都會受到平均每筆交易物品量與Minsup值(出現次數)的變動而有所差異。

而當平均每筆交易物品量改變且Ｍinsup值固定時，在每筆交易當中，擁有較多物品數量則能找到較多的規則，原因在於建立FP Tree時，每筆交易物品越多，樹的深度跟著增加，找到的Pattern Base也較長，能夠產生更多的Frequent Pattern去找到更多規則。

1. Kaggle

將Kaggle搜尋到的資料轉為.arff檔後，用WEKA開啟，將資料進行轉換之後，使用Apriori演算法以及FPgrowth演算法進行分析，下圖和圖為使用兩項演算法分析後得出的結果。

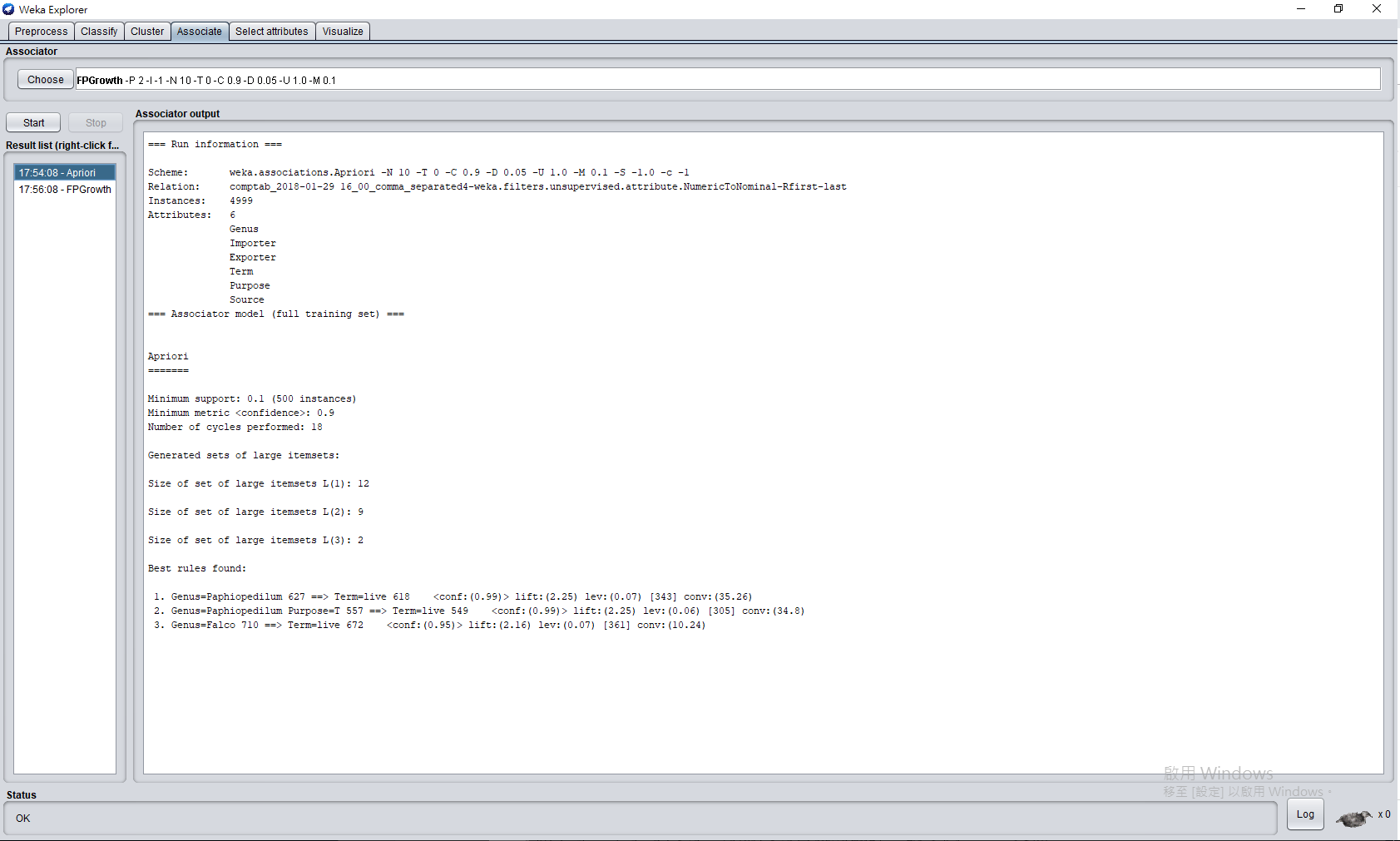


圖　使用Apriori演算法分析的結果

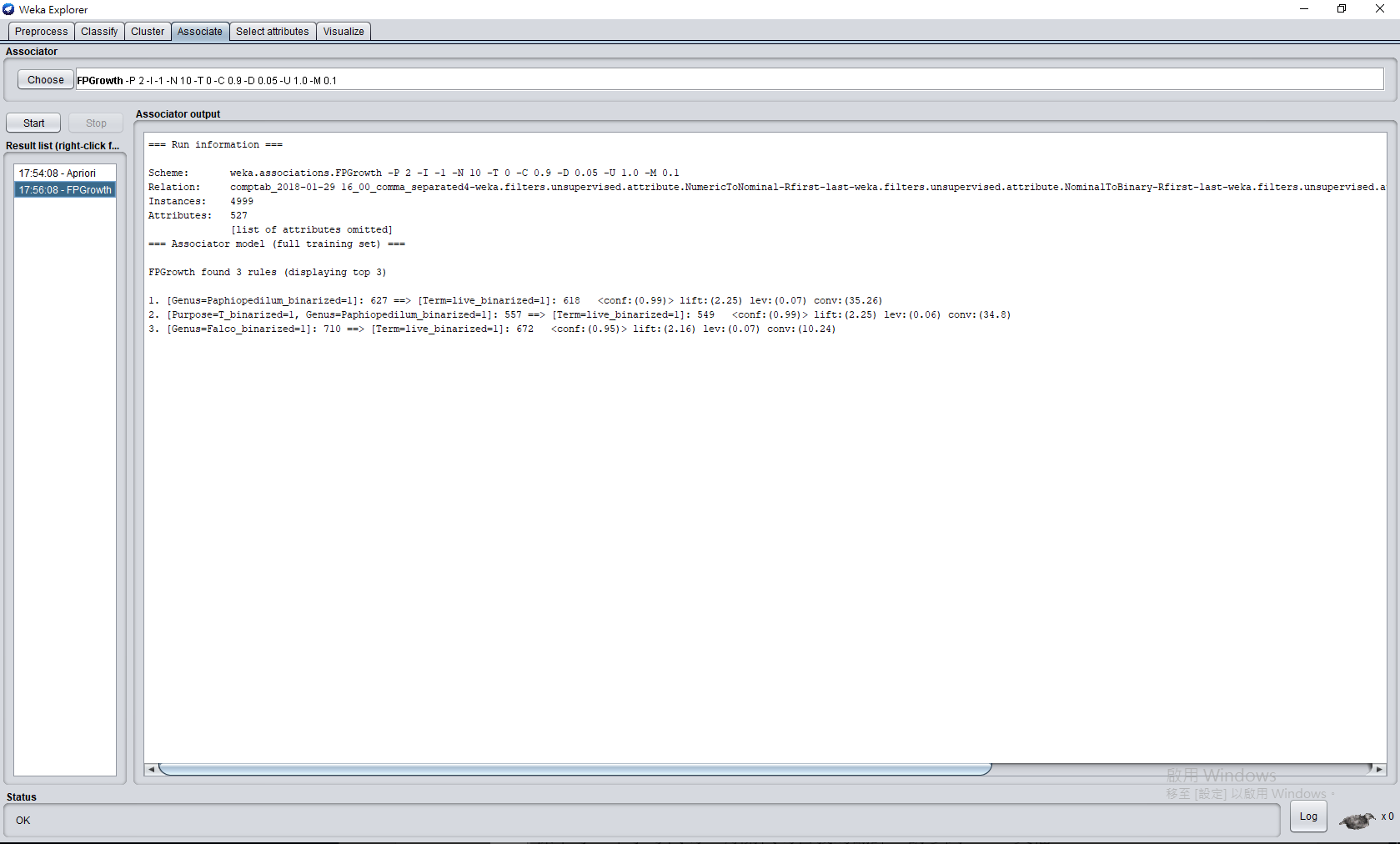


圖　使用FPgrowth演算法分析的結果

使用Apriori演算法以及FPgrowth演算法所分析出來的結果，兩者為一致的，從得出的三項規則可知，兜蘭花（Paphiopedilum）在交易時的型態大部分為活體（live），且用途為商業用途（T代表Commercial）居多，經過網路搜尋兜蘭花的資料，推測其商業用途應該是為各國舉辦花卉博覽會以及栽培所用，且此物種在資料中佔了將近一千筆，可見兜蘭的栽培技術對其保育研究及商業生產意義重大。

而鷹（Falco）交易的型態多半也是活體（live），對尚未取樣的資料進行查找後，發現出口與進口國大部分為中東國家（出口國以阿拉伯聯合酋長國佔多數），而用途大部分為個人用途（P）以及商業用途（T），由此資訊可以聯想到，由於中東國家有馴養獵鷹的傳統，獵鷹捕獲的獵物可以增加沙漠游牧民族的肉食量。如今這項活動還在繼續，且變成中東富人的娛樂，獵鷹大部分會在市場的特殊區域販售，因此獵鷹的買賣被歸類的資料中。

1. 結論

在進行實作的過程中，了解了資料量、Ｍinsup值對於運算時間、規則數、運算成效有極大的影響。撰寫程式碼與WEKA分析的測試結果一致，透過IBM Quest Data Generator產生數個模擬資料，同時用Weka的分析結果與撰寫的程式碼做測試，分析結果成效可能因為資料量大會有所差距，在此測試較大資料集的部分將Ｍinsup值調高，以利快速找到規則並避免程式崩潰。

本次作業在程式部分耗費許多時間，由於個人程式能力不足，發現程式碼會將單項子集也列出為規則，此項小缺失雖不影響結果，但仍需仔細判別。第一次嘗試分析與結果比較，雖然資料不夠完善，但仍可分析出結果，希望未來能夠將此項技術更精進，妥善運用此方法分析其他資料集。