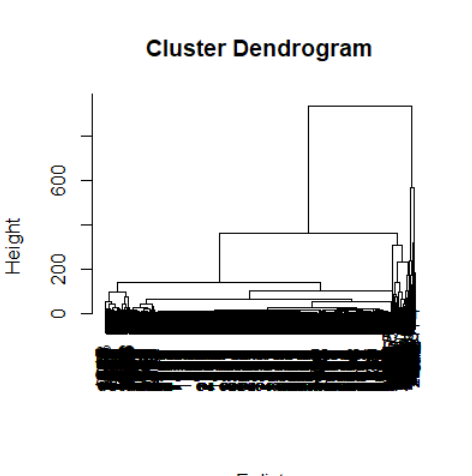
資料探勘FINAL

10427108 何瑞祥

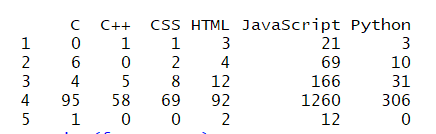
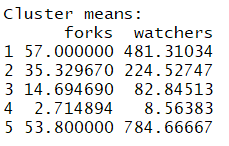
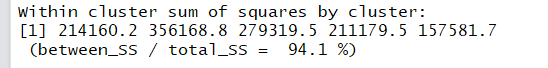
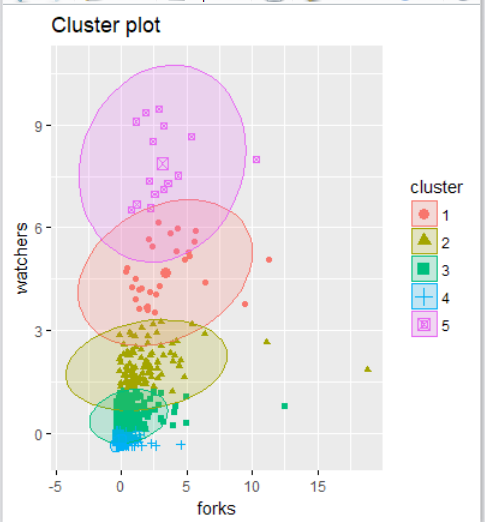
1. 資料:

選用Repos500這個檔案,使用檔案中的forks、watchers、language這3個欄位,forks是代表這個程式碼被其他人複製到倉庫的次數,因為資料太多筆,所以我取樣其中的3000筆,並對JavaScript、CSS、HTML、C、C++、Python這幾種語言做分析­­

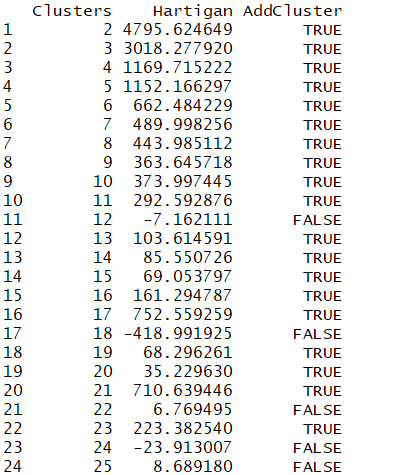
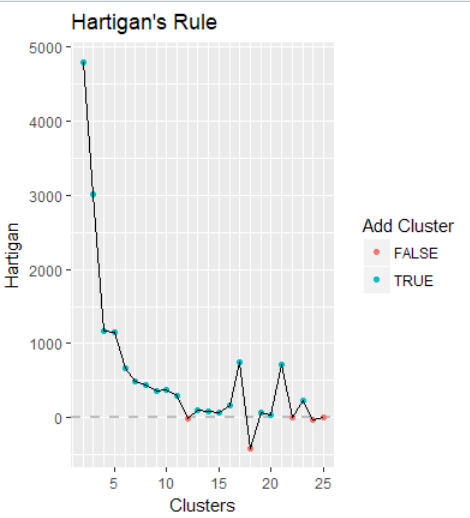
(有把取樣並經過篩選的檔案輸出成.CSV檔,檔名是test.csv) 

2.分群:

一開始用目測的大概是分成5群,其中有一群數量特別多,但是參考價值最低

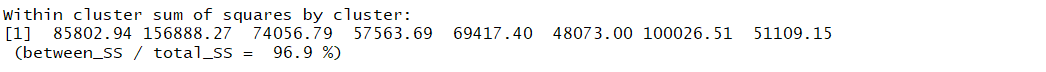
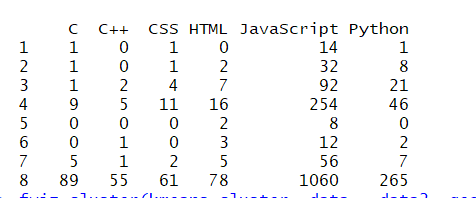
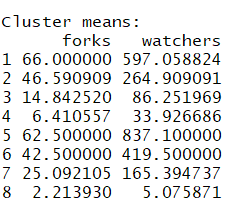
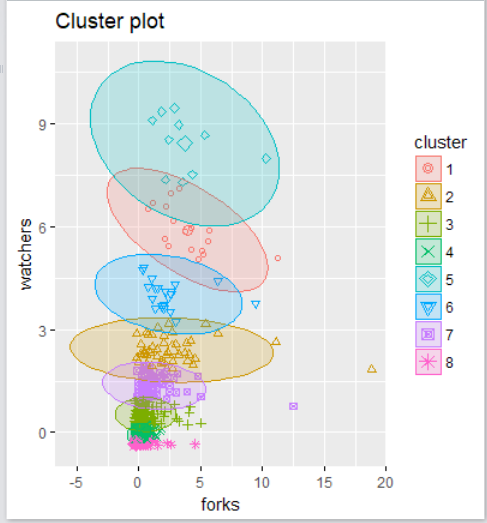
結果:   


使用fitkmeans找K值:

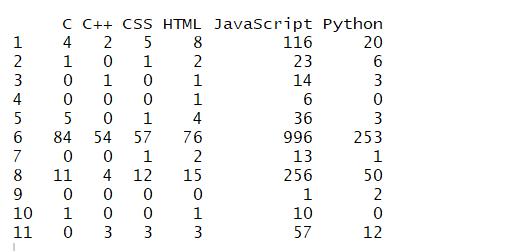


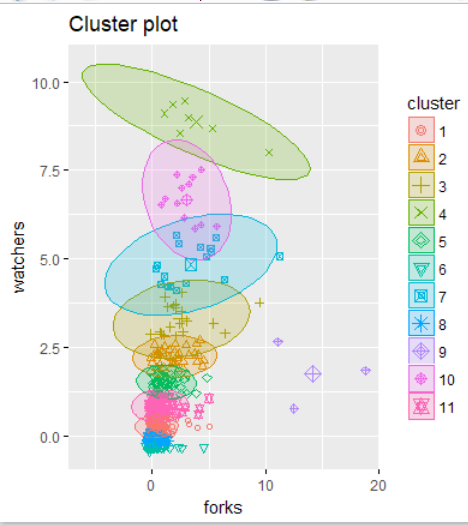
以K=5、8、11做觀察

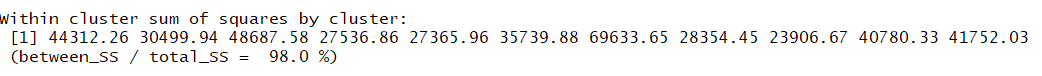
(1)分成8群:

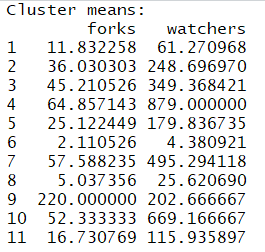


雖然第8群的數量差距很大,但因為forks、watches的cluster means數量太少,所以要看第1、2、5群的趨勢比較準

(2) 分成11群:





第6群是因為forks、watches的cluster means數量太少,所以參考價值不大

,最能看出趨勢的前幾群是第4、7、10群

第一次做的時候,因為沒考慮到cluster means的數量,就直接用數字差異最明 顯的一群做分析,加進cluster means做比較之後的結果是比較可信的

結論:

其實從一開始目測的5群,就可以看出,大概哪一類的語言是比較常被使用,很明顯是跟網頁設計相關的語言最常被使用,可是C和C++,的差別就不太明顯 ,應該是這兩種語言的相似度比較高,沒辦法從不同群的結果看出C和C++的差距;但是以大方向看,不管在哪一群(fork、watcher多),跟網頁設計相關的語言都是最常被使用的,尤其是JavaScript,可能是很多人都是以網頁設計開始接觸程式語言

從forks、watchers的table可以看出,這兩個參數是正相關的,如果是要觀察種趨勢,值越大越有參考價值

心得:

第一次做這個分類的結果很慘,而且方向搞錯了,之後在Github Data的資料檔找到這分資料之後,就有一個比較明確的方向了,做完之後發現用程式對資料做分類,雖然有很多客觀的數據可以看,就像5、8、11群的效率分別是94%、96%、98%,

感覺這學期學的資料探勘課程跟其他的課程真的很不一樣,觀念的重要性遠遠大於實作,因為R真的蠻方便的,所以打程式碼的難度下降許多,經過這學期之後才發現,很多商店的商品搭配不是隨便配一配或是為了清庫存之類的,背後其實有很多理論,才能讓他們順利賣出商品

修正處:

第2次的版本有對fork、watcher做篩選,不能有0或是>1000的值存在

結論的部分是有加入cluster means做比較之後觀察到的情況

第一次因為漏看一個false所以選用K=16,這次的K值都在正確範圍內