**巨量資料分析與應用 期末報告**

基本資料

班級：\_\_\_商資三甲\_\_\_ 分組編號：\_\_\_16\_\_\_\_\_

組長：學號 C109193104 姓名 \_\_\_\_鄭怡芳\_\_\_\_

組員：學號 C109193110 姓名 \_\_\_張珈薰\_\_

題目

桃園市珍貴樹木數據分類

--資料分析**目的**說明--

|  |
| --- |
| 本組利用桃園市珍貴樹木資料2017.csv，將GPS緯度(Latitude)、GPS經度(Longitude)、樹高\_公尺(Tree height)，等欄位分析資料為哪個樹種。 |

資料集描述

資料集個數：1 總筆數：60

資料集1名稱：tree 資料集1筆數：60

檔名1：桃園市珍貴樹木資料2017.csv

來源：https://data.gov.tw/dataset/26129

|  |  |
| --- | --- |
| **欄位名稱** | **欄位描述（包括編碼格式）** |
| **樹木編號(Number)** | 列管序號(String) |
| **樹種(Variety)** | 品種(String) |
| **區域(Area)** | 所在鄉鎮區(String) |
| **地址(Address)** | 所在位址(String) |
| **GPS緯度(Latitude)** | 定位點1(Double) |
| **GPS經度(Longitude)** | 定位點2(Double) |
| **樹高\_公尺(Tree height)** | 高度(Double) |
| **冠幅\_公尺(Crown one)** | 樹徑(Double) |
| **冠幅\_公尺1(Crown two)** | 樹圍(Double) |

分析策略說明

|  |
| --- |
| 說明通過哪些步驟或流程？使用哪些資料集、哪些欄位、用何種方法來分析？  使用多元分類(Multiclass Classification)，   1. 進入Ubuntu作業系統 2. 開啟firefox，輸入資料集網址，下載桃園市珍貴樹木資料2017.csv資料集 3. 將資料集拉進家目錄 4. 建立hdfs資料夾，名稱為tree 5. 將資料集上傳至tree資料夾 6. 進入spark-shell環境 7. 讀入資料 |

程式碼

* 請使用Consolas字型

|  |  |
| --- | --- |
| **指令** | **說明** |
| **進入Ubuntu環境** | |
| start-all.sh | 啟動hadoop，若已經啟動，則不需要再執行此指令 |
| cd ~ | 回到家目錄 |
| wget  https://data.gov.tw/dataset/26129 | 下載鳶尾花資料集 |
| hdfs dfs -mkdir /tree | 在hdfs上建一目錄 |
| hdfs dfs -put ~/Taoyuan\_City\_Precious\_Tree\_Data\_2017.csv/tree | 將資料集上傳hdfs |
| hdfs dfs -ls /tree | 查看hdfs資料 |
| **進入spark-shell環境** | |
| spark-shell | 進入spark-shell |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

HardCopy執行畫面

方法：選好要拷貝的執行結果畫面，按Alt-PrintScreen，出現gnome-screenshot畫面後，點選「複製到剪貼簿(O)」，再到本檔案中貼上。

注意：執行結果超過一個畫面是必然的，所以上下面畫面間的資料應該重複，而不可以缺少。

<<<以下為範例畫面，可以刪除>>>



