

Spark 安裝

1. VirtualBox 安裝
 - a) 無（本來就是 Linux 環境）
2. [Spark 與 Hadoop 安裝](http://simp.ly/p/G5Lpmx)（<http://simp.ly/p/G5Lpmx>）
 - a) 上面的連結為安裝過程中使用到的指令紀錄及說明
 - b) pyspark 執行畫面
 - c) 使用 Jupyter 編輯截圖

[illegible]

(編輯 bashrc，加入以下內容，可在 terminal 輸入 pyspark 時自動開啟 jupyter)

```
export PYSARK_DRIVER_PYTHON="jupyter"
export PYSARK_DRIVER_PYTHON_OPTS="notebook"
```

Jupyter DataMining Last Checkpoint: 11/30/2017 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```

In [2]: from pyspark import sql
        from pyspark.sql.types import *
        from pyspark.sql import functions as F
        from numpy import array
        from pyspark.mllib.regression import LabeledPoint
        from pyspark.mllib.tree import DecisionTree
        from pyspark.mllib.evaluation import MulticlassMetrics

In [3]: read.csv('1061-Data-Mining/HW1_Decision_Tree/character-deaths.csv', header=True)
        # content of the DataFrame to stdout

        # Add Columns
        df = df.withColumn('Death Year', F.when(F.col('Death Year') > 0, 1).otherwise(0).alias('is Dead'))
        df.show()

In [4]: # Generate is Dead columns
        isDead_expr = [F.when(F.col('Death Year') > 0, 1).otherwise(0).alias('is Dead')]
        df = df.select('Allegiances', 'Book Intro Chapter', 'Gender', 'Nobility', 'GoT', 'CoK', 'SoS', 'FfC', 'DwD', * isDead_expr)
        df.show()

```

Allegiances	Book Intro Chapter	Gender	Nobility	GoT	CoK	SoS	FfC	DwD	is_Dead
Lannister	56	1	1	1	1	1	1	0	0
None	49	1	1	0	0	1	0	0	1
House Targaryen	5	1	1	0	0	0	0	1	0
House Greyjoy	20	1	1	0	0	0	0	1	1
Lannister	0	1	1	0	0	1	0	0	0
Baratheon	0	1	1	0	1	1	0	0	0
Night's Watch	21	1	1	1	0	1	1	0	1
None	59	0	1	1	1	1	0	1	1
House Greyjoy	11	1	1	0	1	0	1	0	0
Night's Watch	0	1	0	0	0	1	0	0	0
House Greyjoy	50	1	0	0	1	0	0	0	1
House Targaryen	54	1	0	1	1	1	0	1	0
Night's Watch	18	1	1	0	1	1	0	1	1
None	15	0	0	0	1	0	0	0	0
Arryn	38	1	1	1	0	0	1	0	0
Night's Watch	26	1	0	1	0	0	0	0	0
House Stark	4	1	0	0	1	0	0	0	1
House Tyrell	6	0	1	0	0	1	1	0	0

程式結果

Precision: 52.083333333333336 %
 Recall: 64.1025641025641 %
 Accuracy: 68.51063829787233 %

討論

與作業一相同的前處理（使用 pyspark），但作業一和本次作業在切割訓練及測試資料集時都是使用 RandomSplit 的方式，因此兩次的結果略有不同。