1. Hadoop&Spark安裝：

依據工具書內容安裝Hadoop&Spark(也可以另外上網找資源)，

一般請依照書本的指令，安裝在虛擬機器，如果有足夠的電腦(3台以上)可以安裝實體機

以下網址為該書的指令集合，沒有買書的人建議借書來看會比較清楚

http://hadoopspark.blogspot.tw/search/label/%E6%9C%AC%E6%9B%B8%E6%8C%87%E4%BB%A4%E6%95%B4%E7%90%86

(書本為虛擬機安裝，但該網址並沒有關於虛擬機設定內容，書本有，建議借書來看)

請注意指令可能會有新舊版本與環境問題，因此盡量安裝與課本相同的版本

安裝到書本的第8-4節，輸入spark-shell指令確認是否安裝成功

2. Spark決策樹分析：

本次作業為預測角色是否死亡

資料集在以下網址:

https://www.kaggle.com/mylesoneill/game-of-thrones

其中第二份資料的character-deaths.csv

或下載附件檔案

其中三個欄位 Death Year , Book of Death , Death Chapter 取其中一個欄位當預測目標用即可

請將欄位的空值轉成0(代表存活)，有數值的轉成1(代表死亡)

作業流程 **※限用Spark MLlib支援的決策樹實作**

1)將資料讀取進來(可用pandas套件)

2)資料前處理

2-1把空值以0替代

2-2Death Year , Book of Death , Death Chapter三者取一個，將有數值的轉成1

2-3將Allegiances轉成dummy特徵(底下有幾種分類就會變成幾個特徵，值是0或1，本來的資料集就會再增加約20種特徵)

2-4亂數拆成訓練集(75%)與測試集(25%)

3) 建立決策樹模型

可透過google 搜尋 spark 決策樹 教學，選擇適合自己的教學文章 ex. http://hadoopspark.blogspot.tw/2016/04/spark-mllib.html

作業目標

(1) 產出預測結果　(2) 計算accuracy, recall, precision

3. 作業繳交內容：

(1)程式碼(含註解)

(2)報告(pdf格式)

i. 截圖與步驟描述(安裝的部分將較重要的幾個步驟截圖即可)

ii. 作業目標對應之結果

iii. 討論