**Competition1**

指導教授：李政德　教授

組員:王子豪H24031346 周逸平H24036087

1. **模型介紹(Description)**

我們使用了隨機森林模型(Random Forest)對TF-IDF(term frequency–inverse document frequency)矩陣進行參數篩選，後利用支援向量機(Support Vector Machine)中的Support Vector Regression模型使用篩選後的參數對資料進行分類。

1. 隨機森林模型：

隨機森林是一個包含多個[決策樹](https://zh.wikipedia.org/wiki/%E5%86%B3%E7%AD%96%E6%A0%91)的[分類器](https://zh.wikipedia.org/wiki/%E5%88%86%E7%B1%BB%E9%97%AE%E9%A2%98)，並且其輸出的類別是由個別樹輸出的類別的[眾數](https://zh.wikipedia.org/wiki/%E7%9C%BE%E6%95%B8)而定。其中權重部分所使用的資訊獲利(Information Gain)方法主要利用一資料集的熵值(entropy)與該資料集被某一參數分群時的熵值之差以判斷該參數的重要性。

TF-IDF：

一種用於資訊檢索與文字挖掘的常用加權技術。tf-idf是一種統計方法，用以評估一字詞對於一個檔案集或一個語料庫中的其中一份檔案的重要程度。字詞的重要性隨著它在檔案中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降。

支持向量迴歸模型：

支援向量機模型(Support Vector Machine，SVM)是將例項表示為空間中的點，這樣對映就使得單獨類別的例項被儘可能寬的明顯的間隔分開。然後，將新的例項對映到同一空間，並基於它們落在間隔的哪一側來預測所屬類別。支持向量迴歸模型（Support Vector Regression，SVR）是使用SVM来擬合曲線，做迴歸分析。與分類的輸出是有限個離散的值所不同之處在於，回歸模型的輸出在一定範圍内是連續的。

1. **步驟說明：**
2. 資料前處理：
3. 將training data中的text項與testing data中的text項合併後取tfidf。
4. 將a.取完的tf-idf矩陣分割為training data的tf-idf矩陣與testing data的tf-idf矩陣
5. 參數篩選─隨機森林模型：

將train data之tf-idf矩陣放入隨機森林模型之中，其權重計算方式使用「資訊獲利」進行計算。依此方式將特徵參數減少至2000個。

1. 配適模型─SVR模型：

將2.中選擇的2000個參數放入SVR模型之中，其中Cost設定為10，Gamma設定為2。

1. 預測輸出：

將testing data的tf-idf矩陣放入配適好的SVR模型進行預測，並將預測結果中，小於1的令為1，大於5的令為5，後取四捨五入。

透過以上步驟對測試資料的stars進行預測，最終預測準確度為RMSE得0.8587，ACC得0.4428。

Github url:

<https://github.com/isthereanyusernameNOTtaken/data_sciense/tree/master/report2>