1102\_資料分析與學習基石 個人專題第一次報告

黃振嘉 F74086048

資料集來源figure-eight [Data For Everyone’ website here](https://www.figure-eight.com/data-for-everyone/)

競賽首頁 [Natural Language Processing with Disaster Tweets | Kaggle](https://www.kaggle.com/competitions/nlp-getting-started)

這是一個由figure-eight這間公司製作的資料集，後來被放上Kaggle供大家使用。訓練的檔案包含10875筆資料，分別有keyword、location、text三個標題的內容。此資料集的特色是其中有出現hashtag(#)、at(@)，而且當中也有混雜一些次文化用語如表情符號與縮寫。這些內容與先前處理的標準語法會有一些出入，因此我認為這個資料具有相當的挑戰性。

目的是要訓練出一個模型，用以分辨使用者的Tweets是否與災難相關。我期望自己可以夠熟悉自然語言處理的能力，希望能夠進入排行榜的前百分之三十。