資料科學與社會研究 期中個人報告

經濟三 B05102074 何青儒

第一部份：演講主題摘要

我們身處的二十一世紀是一個充滿資料（Data）的年代。

對一般人而言，科技的創新與進步讓接觸網際網路的成本愈來愈低；而對廠商而言，儲存裝置也變得愈來愈便宜，人們開始意識到過去覺得是「佔空間」的雜訊資料其實富有潛力。例如裝置上的搜尋紀錄、瀏覽紀錄、網站停留時間、透過何種裝置瀏覽、在何時何地造訪等等，而如何將這些零碎的、大量的、未經整理的「原始資料」轉變成有用的、易被觀察的、系統性的「富有價值的資料」就是資料科學（Data Science）這門學問的主要目標。

東京國際大學（TIU）的陳釗而教授在第十一週時以 *Data Science and Economics* 為主題，和我們解釋「資料科學」和「經濟學」之間的關係，就是想要告訴在座的各位，即使資料科學更常被視為是結合電腦科學（Computer Science）、數學與統計學（Math and Statistics）、商業知識（Domians/Business Knowledge）的學科，和經濟學沒有什麼直接的關聯，但經濟學家還是可以透過長期以來被訓練的經濟直覺，去嘗試解釋資料與資料之間的關係。而如何將經濟學的思想應用在資料科學之上，就是我們在未來該去思考的問題了。

第二部份：研究主題草圖

定義研究問題

在 Assignment#6 颱風新聞的切字練習時，我在作業最後面提到我原本想從練習中探討的問題：聯合報的新聞或是社論是否有媒體政黨自助餐的傾向？意即「520 分水嶺」的現象？

所需資料變數

- 新聞／社論的原始文檔
- 新聞／社論發佈的時間（區分各個時間點）
- 新聞／社論的作者（區分同一個作者在不同時間點會不會有相差甚遠的想法）

分析方法

在這裡，會用到 SnowNLP 這個套件來處理中文文章的切字和情感分析。

> SnowNLP 的 Github 專頁：https://github.com/isnowfy/snownlp

可能的結果

在 Assignment#6 的後記裡，我預期因聯合報在政治光譜上屬於很藍很藍的那邊，所以在民進黨執政期間發生的災害或意外，一定都會說是中央政府的問題（如「中央治水預算分配不均」）；而國民黨政府期間則會推說是地方政府的錯（如「地方治水設施久未更新」）。

但原本僅透過切字去分析不同時期出現「中央政府」、「地方政府」的頻率，卻沒辦法得出顯著的差異，這可能歸咎於「切字分析」只採用「出現了」幾次的頻率當作分類依據，而沒有考慮到「中央」或是「地方」政府的新聞內的「態度」可能會不一樣。有可能這篇出現的「中央」，是在怪罪中央政府沒有編列治水經費的缺失，但另一篇卻是用來稱讚「中央」在事後回復家園的時候處理的很好；這篇出現的「地方」可能是在說地方政府都沒有協助救災，但另一篇有可能是在說「地方」在這次災後有幫忙協助重建。我們應該還要考量到每一篇新聞是用持平、稱讚或是譴責的語氣去做分類（即情感分析），再從中取出「中央」和「地方」的差別。

所以，我打算主要先把所有的新聞或是社論依據時間和新聞主要提到的對象分類成兩類：

- 依期間分類：
    - 民進黨政府執政期間
    - 國民黨政府執政期間
- 依新聞主要提到的對象分類：
    - 中央政府
    - 地方政府

然後再透過情感分析，去分類在不同時間點、新聞對象，新聞或是社論對其所產生的「態度」之分。例如在民進黨政府執政期間，「地方政府」的新聞態度可能傾向於稱讚，而「中央政府」的新聞態度可能傾向於譴責；而國民黨政府執政期間則是相反過來。使用了 SnowNLP 的情感分析可以讓我們從中拆分兩種「情感態度」，從原本只區分「中央」和「地方」擴展到「稱讚中央／地方的」和「譴責中央／地方的」。

但是在這裡要處理的是，如果使用 SnowNLP 這類第三方、別人已經開發的文字情感分析工具，他們大多的訓練模型是在原作者所使用的原始文本數據（如 SnowNLP 的原作者在 Github 的 Readme 裡提到他是使用「買賣東西時的評價」訓練出來的），未必適合我們想對「新聞／社評」的應用情況。

如果要增加分析的準確率，我們可能還是得訓練自己的模型，例如透過 scikit-learn 去做 ML，但這樣又會出現一個尷尬的情況：如果我們今天是訓練一個評分系統，在這個系統裡面有 rating（可能是一到五顆星）和使用者的 comment，那我們就可以透過 rating 去訓練這個 comment 是屬於稱讚（例如四顆星或是五顆星時的 comment 是長怎樣）或是譴責（例如一顆星或是兩顆星時的 comment 又是長怎樣），但新聞呢？一篇原始的新聞文本是沒有 rating 這個機制的，我想到的一種旁門左道的方法就是透過計算這篇新聞在的 PTT 八卦、政黑、地方板的推噓數當作 rating 的參考值，但又會出現另一個問題：PTT 上的使用者其實也是會有政黨自助餐的現象，我們單單只用推噓「總數」去計算，真的能夠區分出是稱讚還是譴責嗎？