

Text mining流程：

一.爬蟲

- 1.看板studyabroad

- 2.抓到標題有”錄取”的

- (1.)“心得”跟”錄取”選擇了錄取，因為錄取的文章數量較多，大概有2000多篇

- (2.)”錄取”文章的格式內容較固定。

- 3.GRE的分數有經過改制(2011-08-01)，所以將爬到的文檔分成改制前以及改制後，先處理改制後的文檔。

- 4.將文檔以文章id為檔名。

二.正規

- 1.目標分數 GPA, GRE, TOFEL 三個量化數值

- 2.原本做法：用正規表示式去找 GPA, GRE, TOFEL三個Pattern，然後抓出Pattern後面的數字。

- (1)問題1.：最好情況 GRE:332，但是有許多情況是『GRE：\n332』，就抓不到了。

- (2)問題2.：這種方法很難將東西取得乾淨，而且取完Pattern那行後，還要再去抓特定值，很麻煩。

- 3.後來做法：GPA的範圍2.4~4.3、GRE的範圍290~340、TOFEL的範圍80~120，因為值的範圍皆沒有重疊，只需用正規表示式去抓範圍內的值就可以了，變的精準度相當高。

- (1)問題1.：GPA大部分的人表示方法是(4.12/4.3) or (4.00/4.00)，所以會抓到重複的數字，解決方法是，同樣都抓出來，但取較低的當作本篇的GPA

- (2)問題2.：GRE跟TOFEL都有人將分數分開列出，而非直接列出總和，故範圍必須再變，且必須去使用python一個檔案一個檔案去處理，而不能使用Linux下的grep指令直接抓所有檔案的GPA、GRE、TOFEL然後再pipeline到另一個檔案內。

(3)問題3.：推文的重複敘述，在推文中可能會有覆述的狀況出現，解決方案為將推文列表全部刪掉。

(4)問題4.：標題會有時間序列，會導致正規表示式所抓到的東西不好，解決方案為將標題相關序列刪掉。

4.無法解決之問題：有人考了多次的GRE、TOFEL並列出，因此可能一篇有多個成績，需人工校正。

Clustering流程：

零、希望能夠依照所得特徵來將學生作分群，並得出結論

一、找出最好的群數量

- 1.SSE - 2
- 2.Silhouette - 2
- 3.Gap stat - 5
- 4.最後選擇4來當群的數量

二、用不同的方法去分群

- 1.simple kmeans
- 2.dist cluster method

三、分群所得圖，群聚關係並不明顯，因此試著挑一些特徵出去，讓分群效果更清楚

- 1.AIC - Forward & Backward
- 2.BIC - Forward & Backward
- 3.最重要的特徵為 GRE、GPA、TOFEL、LOR，其中以GRE跟GPA所得之圖分群效果最好