IR 期末專題書面報告

第九組、Twitter 使用者特性分析器

R02725002 何鎮濠、R02725048 許宏瑋

一、 動機與發想

原初構想

我們在尋找合適的 IR 應用題目時，其實一開始是注意到 twitter 使用情境上的特性：twitter 它的使用者行為、tweet 之內容都是相當程度地公開的！這有利於我們取得與應用其上的資料來做 IR 分析。

其實 twitter 的這個特性，源自於其服務的核心構想：創建一個大家都可以公開發聲的平台（且所發的內容必須極簡短易讀）；這點我們也可以在其中主要活動的動詞是＂tweet 啼叫＂、和其服務的 icon 是一隻啼叫的小鳥上，稍微能窺知一二。

也因如此，twitter 其實並沒有明確的朋友、社團、子團體的概念在裡面（顯然是故意的），使用者依序接收到的各個所追蹤之 twitter 發的短文，它們都是平起平坐、先來後到，它們都按照短訊發出的時間被排列、被擠下去。

使用者在瀏覽的時候，如果它追蹤的人到達一個量、它就會發現 twitter 的訊息版面頗為繁複、更新迅速，配合上各種可能的使用者頭像，整個版面視覺上就有點凌亂、難以一目瞭然。

所以我們想要對於快速、較為密集而且多樣化的 twitter 簡訊，進行一個恰當的分類、輔助閱讀；而且設計上不希望破壞到 twitter 原本的使用者經驗。

可能的方法、我們初步構想是在 twitter 的簡訊格子的最前端、或周圍加上一個顏色的示意標記、像是最近一些 email 集成之 client 端 app 一樣，有一個代表不同分類的長條狀色塊，協助使用者更容易一看就抓住視覺上的重點。

這樣我們把我的題目命名為：社群網路服務之信息分類與排序（排序是也許我們能夠做的延伸，故收錄在題目裡，主要我們是想要藉由分類來協助使用者的閱讀體驗。）題目之餘，我們也在 proposal 當中設想了這個實做的一些商業亦或是社會科學研究上延伸應用的可能性。

潛在難題、抉擇與評估

在後續的評估當中，我們發現我們的構想在實做上會有一些問題，其中最大的問題就是：其實我們的構想，是想要做一個短信息的分類器；但當我們要分類的信息很短的時候、分類器的效能未必很好。

我們的實作時間有限，可是我們仍希望能有明確的成果！我們不太願意投入一個不確定

成效為何的項目。這是一個課程的實做練習，雖然也可以看成是一個做實驗的機會，但這仍不是我們正規的實驗題目。相較於進行一個實驗，我們更把它看成是一個雛型元件的實踐，它必須要有更篤定的功能面。

所以我們做了調整：一個 tweet 的篇幅可能太短，我們就把相同人的數十篇、甚至數百篇 tweets 集合在一起，這樣就可以做不同使用者之發文風格的分類！如果我們的分類項目選定得當，這個做法結果應該會很有趣！

並且，我們想盡可能地專注在重要的核心功能上，我們就把使用者的風格先分成三類，以求實做上的明確，並且我們目前只想要處理英文的部分、這會讓我們在方法的構築上更加精確。三個分類分別為：情緒化發文者、商業化娛樂發文者、政治時事言論的發文者。

有了 twitter 使用者的發文風格的資料，也許能夠有許多其它延伸的商業應用，比如說針對使用者的發文風格去回溯這個使用者的喜好，進行精準行銷等等。

但比起行銷上的應用，我們認為如果我們的分析器做得好、並且延續我們原初在使用者閱讀上的構想，我們新的設計仍然能夠協助改善使用者的閱讀體驗。並且當使用者對自己的發文進行風格分析的時候，其實也能讓使用者有個多了解自己過去發文習慣的機會，別有一番趣味。

二、　方法構築

方法設計

決定了三個特色明確的使用者風格分類後，我們想要用分類的方法為使用者的發文風格做分析。把一個使用者近期的 tweets 接在一起、看成是一個文件，計算該文件被分入三個分類個別的適合值，適合值最高的那一類即為這個使用者最突出的發文風格；並且我們仍然提供其它兩個分類的數值，讓使用者能參考自己在其它屬性上的表現為何。

好消息是，我們有一個相近的程式碼、作業三！我們如果能確保在 Training Data 上蒐集的文章風格確實能代表我們理想中所要分出的那三類使用者發文風格，我們就已經成功一大半了！考慮到 twitter 使用者的多元化，我們需要有多元的 Oracle 文章來源。

在政治時事上新聞、官方人員的發言、和熱中時事談論之徒的文章會是好的選擇。而商業化娛樂的部分，各種廣義上商品的廣告、使用者體驗分享等等，都會是我們的目標收錄對象！比如說音樂評論、電影評論、商品開箱文、企業的廣告、商品發表會、商品廣告等等都是。另外，像是談飲食文化的文章、食記、遊記、旅遊介紹、風景名勝的摘要、廣告等等，也不可或缺。情緒化發言的部分，我們就會需要一些更為口語的詞彙，我們應該集中到各大論壇和 twitter 上的使用者本身來蒐集。

其實最重要的，就是要組成一個合適、貼近我們理想分類目標的 Oracle。我們應該在每一類都截取相關有代表性的 twitter 使用者的發文集。

以下，就讓我們來看看我們是怎麼實做的！

實做摘要

。開發環境：Java IDE、Twitter API、IR_PA

。實作流程：先從網路上 BBC 新聞、飲食/旅遊 Blog、 購物網站、電影評論、網路論壇，以及透過 Twitter API 爬取各類型 twitter 使用者各 150 篇 tweets 所取得的總合 47 筆 Oracle Data,使用 Naive Bayes 計算分類機率值,配合 Likelihood feature selection 方法取出 2500 個 terms 為計算依據，並建構成系統的 Oracle Dictionary，最後從 Twitter 上爬取 24 名使用者各 150 篇最近發佈的訊息作為 Testing data。讓系統針對預設的 3 大類別對 20 名 Twitter 進行 Classification。

Emotions Terms Score

6073	
term	likelihood
fuck	10.222507
lol	10.222507
tire	10.222507
ass	8.8892632
hot	7.8948545
gotta	7.4297271
heat	7.4297271
omg	7.4297271
bitch	7.4297271
http	6.4209619
gonna	6.2506032
shit	6.2506032
babi	6.2506032
miami	6.2506032
tonight	6.2506032
tweet	5.3927789
mad	5.3927789
leg	5.107758
hoe	5.107758
idk	5.107758
dude	5.107758

Entertains Terms Score

6073	
term	likelihood
econom	6.7301822
enjoi	5.2830305
peopl	5.1470399
said	5.0164227
presid	4.846952
economi	4.5836234
featur	4.4307213
leader	4.1845317
polit	3.9342704
inspir	3.6298959
best	3.413672
littl	3.3722425
budget	3.3155653
obama	3.3155653
unemploy	3.3155653
govern	2.9794545
version	2.8798337
fresh	2.8798337
cover	2.8798337
produc	2.8798337
award	2.8798337

Politics Terms Score

6073	
term	likelihood
best	10.899882
econom	8.1771965
littl	7.1009526
presid	6.5036883
love	5.5518031
economi	5.2830305
season	5.2648158
eat	5.2648158
make	5.0858645
watch	5.0716934
state	5.035603
lot	4.846952
hour	4.5836234
http	4.5836234
enjoi	4.5836234
minut	4.5836234
photo	4.5836234
polit	4.4307213
feel	4.3666067
star	4.1845317
complet	4.1845317

三、 執行回顧

經過系統對 24 筆 Testing data 的分類過程後，結果輸出如下列圖表所示，顏色最深者為分數最高的所屬分類，反的則最低：

Twitter	Emotions	Entertains	Politics
EmmaWatson	-899.367	-881.702	-989.242
IanMcKellen	-822.77	-746.602	-889.121
LadyGaga	-1321.07	-1256.94	-1461.88
LeonardoDiCaprio	-1164.35	-1051.71	-1131.51
TaylorSwift	-939.702	-837.896	-1015.03
TomHiddleston	-1214.84	-1132.27	-1275.83

知名藝人 6 筆

Twitter	Emotions	Entertains	Politics
i_McDonald's	-719.167	-659.66	-808.452
i_SUBWAYRestaurants	-442.396	-424.531	-513.965
i_TeslaMotors	-627.879	-584.488	-671.767
i_TheWhiteHouse	-1274.27	-1160.13	-1105.51
i_ToyotaUSA	-1129.53	-1048.65	-1160.03

商業與政治組織 5 筆

Twitter	Emotions	Entertains	Politics
p_HarryReid	-1500.86	-1409.36	-1273.11
p_JeffSessions	-1114.04	-1039.17	-910.638
p_JohnMcCain	-1450.06	-1311.93	-1321.75
p_RandPaul	-1368.28	-1299.26	-1164.91
p_TedCruz	-1496.23	-1424.00	-1285.00

政治人物 5 名

Twitter	Emotions	Entertains	Politics
c_BroncosDoe	-974.603	-1119.99	-1245.61
c_ClaytonBaker	-1039.6	-1024.57	-1186.42
c_EmilyAkridge	-864.685	-971.03	-1074.54
c_JohnWagner	-1288.05	-1204.87	-1057.94
c_Julio	-1631.53	-2363.93	-2506.89
c_Niyke	-1082.73	-1533.54	-1629.41
c_Jana	-581.308	-641.64	-744.671
c_FakeJohnKennedy	-1070.22	-989.007	-975.195

一般民眾 8 筆

從上列圖表可以看到知名藝人與商業組織都被分到 Entertains 一類，而白宮(theWhiteHouse)則被分到 Politics，與一般認知和預期結果非常吻合；對於政治人物群，除

了 JohnMcCain 以外都被分到預期的 Politics 一類中，但 Entertains 與 Politics 兩類的分數差異卻非常微小，經過後續的 Twitter 訊息查探後發現 JohnMcCain 議員最近因女兒出嫁而發佈和分享了較多與個人相關的 twit 訊息，因此影響了最後的分類結果；而為了測試系統對於 Emotions 一類的分類表現，因此在一般民眾的分類部分按照使用者所發佈的 Twitter 內容挑選了較多預期屬於 Emotions 一類的使用者作為主要 Testing data，副以 ClaytonBaker 作為 Entertains、JohnWagner 與 FakeJohnKennedy 作為 Politics 的預期分類候選人，而最後的輸出也顯示了這 8 筆 testing data 的分類結果都符合了當初的預期。

四、　其它可能性之盤點

綜合上述的執行結果，可以看出分類系統具有良好的分類水準，然而還有不少的完善與發展空間，例如系統目前只能分析英文 twitter，以及在處理上沒有考慮特殊符號；分類結果並沒有標準化，而且缺乏一個友善的圖形化使用者介面，因此在結果的可讀性上相對較差。

除此以外，系統還可以加入諸如廢文指數等各類分析指標，讓系統的分析內容更為豐富與有趣；甚至可以根據系統的分類結果為 Twitter 的企業或名人使用者推荐潛在 Follower。我們與老師在這次的報告之後，都認為這次的實做是一個有趣的實做！