ShrimpNews 蝦新聞

動機

看新聞報導是生活的一部分，但很多人會覺得台灣新聞媒體素質低落，有時看到一些新聞會讓讀者覺得『這也算新聞』或『記者真好當』等等讓人嘆氣搖頭的『蝦新聞』。我們希望能發展出一套系統，能從各種新聞來源爬取新聞，有效辨識出其中的「蝦新聞」，將之陳列出來，作為大眾紓解情緒的管道、茶餘飯後的話題，也希望遭列出的新聞記者能坦然面對、記取教訓、堅持改革、大步向前。

解法

我們觀察發現，對於一篇讓讀者感到莫名其妙且不必要存在之新聞，容易收到某些有相同特點的評論。我們相信垃圾新聞的共同點在於，讀者群看見這些垃圾新聞，會有相類似的評論和留言。因此，分析每則新聞的評論，我們想從這些評論中找出特徵，進而從一篇新聞的評論分析此篇新聞是否『蝦』。

資料來源與標記

為了在爬取新聞時，也能一併取得群眾對新聞的評論，我們選擇的資料來源為今日新聞網 NOWNews、蘋果新聞 AppleDaily（4700 則）以及 PTT 八卦板（3894 則）。前二者的新聞頁面裡有 Facebook 外掛留言元件，能以 Python Scrapy engine 將新聞內文與 Facebook 的群眾留言一網打盡。然而，PTT 八卦板並無 Web 版本頁面，故我們以 C++ agent 抓取轉錄到 PTT 八卦版的新聞，並且把下面的推噓文都當作本則新聞的評論。

每一篇以上述方式搜集到的新聞，都需要一個『蝦』或『不蝦』的 label。我們架設了一個網頁，並在社群網路上請朋友、親戚以及 NTU 板上的熱情同學們以人工方式標記 label。使用者看完一篇新聞之後，可以選擇『蝦』與『不蝦』；選擇『我不知道』即是此題跳過不列入統計。每一篇新聞都由多於三個人標記，並且採取多數決來判定此篇新聞最後的 label。

在標記之前，我們告知使用者我們對蝦新聞的定義：

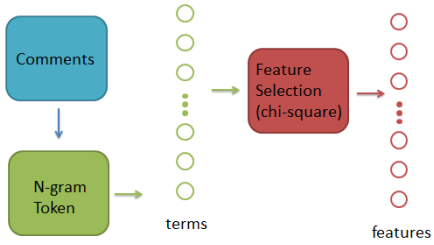
1. 讓讀者感覺此篇新聞沒有存在之必要，讓人有"這樣也一篇"之感
2. 新聞內容錯誤或是荒謬，讓人覺得記者沒搞清楚狀況未求證，素質低落待加強。
3. 任何讓你感到"這好蝦!" 的新聞

並且提供兩則蝦新聞示例。使用者進入標記系統後，除了新聞內容之外，也會同時看到此篇新聞有哪些評論。去除沒有評論、無法進行標記的新聞，最後我們得到了 4725 篇有標記的新聞作為訓練資料，其中有 849 篇 NOWNews 新聞、200 多篇 AppleDaily 之新聞，以及 3600 多篇 PTT 八卦板之新聞。下圖是使用者所見介面，含新聞內容、評論。

系統流程

1. Feature Term Selection

經過上述的標記流程，每則新聞現在都有各自的評論(comments)和此篇新聞是否為『蝦』的 label。



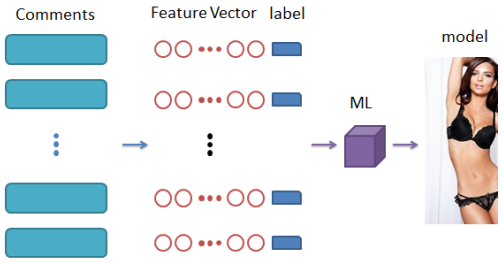
我們希望 feature 能夠捕捉大眾在評論蝦新聞時，所使用的共通詞彙，但使用任何字典式的斷詞工具，應該沒有辦法反應大眾在網路上的詞彙使用習慣。因此，我們使用 N-gram[1] 將 comment 斷成上億個 N-gram term（N=2~8），而後使用 χ-square statistic 來進行 Feature selection。為了將「和蝦 label 最有關」和「最無關」的 term 分開，故修改 χ-square statistic ：

$$\chi^2(t,c) = \frac{\text{sign}[AD - CB] \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

我們選擇取和「蝦」最有關的 2000 個 term、加上最無關的 2000 個 term，得到 4000 個 feature terms。

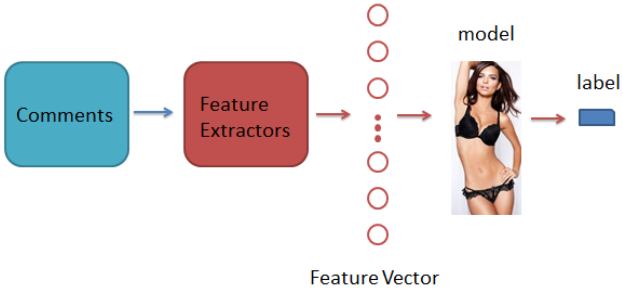
2. Training Process

我們計算一篇新聞的評論裡，各個 feature term 出現的 term frequency，組成此篇新聞的 feature vector（4000 維）。4725 篇標記過後之新聞，以下圖所示之方式來訓練分類器 model。



3. Testing Process

訓練 model 的一週後，我們另外再從今日新聞網、蘋果新聞與 ptt 八卦板爬取 6/12、6/13 之新聞（共 3123 篇新聞，各新聞網所佔比例和訓練資料接近）作為測試資料。我們將其評論根據前述的 feature terms 轉成 feature vector 之後，丟到 train 好的 model 中，便可得到此新聞的 label。



機器學習演算法之比較

我們用 Weka 比較了三種分類演算法的表現：LIBSVM、Naïve Bayes 和 KNN(K=1、3)，分別作為 Discriminative model、generative model 與 Retrieval-based model 的代表。其中 KNN 之 K 值僅選用 1 與 3 ，是因為標記為蝦的資料相當稀疏，資料庫中大多為標記不蝦的新聞，若 K 值選得太大，選出的「neighbor」可能大多會是「不蝦」的新聞，導致 false negative 過高。KNN 使用 Euclidean distance 作為 distance metric，並以距離的倒數做 weighting。SVM 方面，我們使用 grid search 找到最合適的 cost 與 gamma，使用的 kernel function 為 RBF。下面是 5-fold cross validation 的結果。

	TP	FP	FN	TN	Precision	Recall	F1
LIBSVM	36	15	128	4546	0.71	0.21	0.33
Naive Bayes	124	2032	40	4546	0.057	0.756	0.106
1NN	18	49	146	4512	0.269	0.110	0.156
3NN	2	0	162	4561	1	0.012	0.024

從 False positive 之數量，不難發現 Naïve Bayes 看到黑影就開槍，很多都猜是『蝦』，導致 precision 相當低。3NN 和 1NN 相比，classifier 判定是 True 的比例大大下降，變得非常的保守，也因為 false negative 很高，拉低了 precision，一如之前的猜想。最後，LIBSVM 是三種 model 裡面表現最好的，只是其行為也趨向保守，recall 僅 0.21。

修正

上述結果實在不如預期。我們檢視使用者標記的結果，取出來的 feature term 看起來也和我們認知上，大眾會下的評論有所差異。或許是因為在標記時，我們沒有提供「不蝦」新聞的反例吧，有很多新聞在報導上有盡到客觀陳述事實的責任，但大家會因為所報導的對象和自己立場不同，認為被報導的人行為很瞎，就把這篇新聞標記成「蝦」。另外，有另一批重複度高的花邊新聞，讓我們組員一致認為沒有存在價值，但網友們卻都把這些新聞標記為「不蝦」。

為了讓 ground truth 更接近我們所認知的「蝦新聞」，我們(1)人工剔除了使用者標為「蝦」但我們不覺得蝦的新聞。我們也(2)決定自行取出下表中的 term 當作 feature，將評論中有含任一個 term 的新聞都標記為「蝦」。最後，聯集(1)(2)兩種方法所標記出的蝦新聞，作為最後的 ground truth label。我們將這份新的 label 重複一次前述的步驟，希望 feature selection 的結果可以產生新的、不在下表的 feature term。

妓者	妓者不意外	旺報不意外	2 沒壞	Cd
頭噓	新聞專業	甘我屁事	費文	廢文
干我屁事	關我屁事	這也能當新聞	e04	爛新聞
沒營養	垃圾新聞	中天不意外	女支	

值得高興的是，Feature selection 過程除了選出上表我們所人工選出之 feature term 之外，也挑出了其他較有代表性的 feature term，如下表所示：

記者	沒營養	媒體	這種
的記者	版面	出來	媽的
無聊	曝光	Po	你娘沒營養的

我們用這樣的資料重新訓練分類器 model。此時各種演算法之 F1 score 都有所提昇，SVM 的表現也仍然是最好的。

	F1-score
LIBSVM	0.677
Naive Bayes	0.298
1NN	0.358
3NN	0.154

Demo Site

我們將上述之測試結果放在下面網站：
http://shrimpnews.herokuapp.com/

站上收錄著 6 月 12 和 13 日（presentation 前兩天）在 AppleDaily 和轉錄到 PTT 八卦版的蝦新聞，以 LIBSVM model 來進行「蝦」與「不蝦」的標記。

從網站列出的標題，我們發現呈現的新聞差不多也是我們認知上的『蝦新聞』：

暴紅「煎包妹」 山崩照躍頭版
「范范」范瑋琪與「黑人」陳建州 - 代言 Forevermark
智慧局中性、善良　官員盼鄉民別再「制裁制裁」喊不停
602 強震臉書遭地震文洗板！正妹怒：滑手機比命重要？
〈獨家〉點 5 分煎 3 分　鐵板 200 度持續加熱
范范天涯小歌女 信仰抗憂鬱
范范誇黑人壯如牛 每天要努力做人

Future Work

若我們將新聞分類，依照不同類型的新聞分別取 Feature，依照不同種類的新聞下去判斷，應能提昇分類器的表現。

工作分配

周世恩：寫 crawler 爬 NOWnews、AppleDaily 之新聞資料、人工標記、發送 ptt 幣酬謝標記人員。
詹雨謙：寫 telnet crawler 爬取 PTT 八卦板資料、人工標記。
梁翔勝：整合上述新聞與資料標記，實作 feature selection 並進行機器學習，放入 demo site 之資料庫。改進書面報告。
吳潔薇：寫人工標記網頁與 demo site、人工標記以及整理、準備報告投影片、上台報告、書面報告撰寫整理。

Reference

[1] N-gram 斷詞工具使用 https://github.com/timdream/wordfreq，此工具支援中文、英文、日文，且會自動排除 stopword 以及所有標點符號。