當情意席捲而來－華語流行音樂歌詞分析

彭俊穎 b97705023
鍾欣廷 r00725009
王維新 r00725041
魏良丞 r00725052

1. 動機

生活中常常有想抒發心情或感想的時候，如果這時發生找不到適當詞彙來紀錄的狀況，是一件相當令人困擾的事情，因此本專題的目的為自動產生符合使用者心情的字句以適度表達使用者的狀態。在此，我們試圖利用華語流行音樂的歌詞，來達成以上目標。透過專業作詞人的文字，常常可以使聽者產生共鳴，甚至有讓聽者覺得該歌詞是為其量身訂做的經驗。

因此，我們將利用華語流行音樂之歌詞進行文字探勘的分析，並進行以下應用：一、歌詞性質之分類；二、作詞人用詞差異之分析；三、不同年代間情歌用詞差異之分析；最後利用以上結果所產生的五百個最能代表情歌的詞彙，做成自動歌詞產生器。

2. 資料處理

資料處理總共可分為資料收集以及中文斷詞兩大類。資料收集由全體組員共同完成，從 1981 年至 2012 年共收集了 581 首華語流行音樂之歌詞，而後將所收集的資料透過中央研究院所發表的中文斷詞系統進行斷詞作業。

由於所有的分析方式都需要將原始語料斷句並斷詞後，才能處理分析，我們取得的原始語料本身已使用標點符號或空白符號斷句，但是並沒有處理斷詞。我們所取用的歌詞資料是由中文組成，因此若要處理原始語料的斷句，則必須對中文語料進行斷詞處理，「中文語料的斷詞處理」本身就是一個獨立的研究主題，我們這次的研究焦點是對歌詞用法的分析，因此我們並不花太多人力處理斷詞問題，而是向中研院資訊科學研究所申請使用「中文斷詞系統」，並在申請成功後撰寫 PHP 工具，對已收集的語料進行預先斷詞處理，並針對使用者的輸入（分類測試）資料即時處理斷詞以便後續分析之用。

「中文斷詞系統」傳回的資料包含將原始語句斷詞後的結果，以及各詞組可能標記的詞性，因為我們後續的使用很少用到詞性標記的結果，加上「中文斷詞系統」標記的詞性較為細碎（如 Table 1，來源: 中研院平衡語料庫詞類標記集），因此我們將部分詞類合併（如形容詞、名詞、動詞等大類），並將可能為 stop word 的詞類（如: 感嘆詞、標點符號、「的」、「是」…等）結果刪除。

Table 1

精簡詞類	簡化標記	對應的 CKIP 詞類標記[1]	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如：和、跟*/
POST	Cab	Cab	/*連接詞，如：等等*/
POST	Cba	Cbab	/*連接詞，如：的話*/
C	Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
ADV	Da	Daa	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Neu	Neu	/*數詞定詞*/.
DET	Nes	Nes	/*特指定詞*/
DET	Nep	Nep	/*指代定詞*/
DET	Neqa	Neqa	/*數量定詞*/

POST	Neqb	Neqb	/*後置數量定詞*/
M	Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
POST	Ng	Ng	/*後置詞*/
N	Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
Nv	Nv	Nv1, Nv2, Nv3, Nv4	/*名物化動詞*/
T	I	I	/*感嘆詞*/
P	P	P*	/*介詞*/
T	T	Ta, Tb, Tc, Td	/*語助詞*/
Vi	VA	VA11, 12, 13, VA3, VA4	/*動作不及物動詞*/
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11, 12, VB2	/*動作類及物動詞*/
Vt	VC	VC2, VC31, 32, 33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方賓語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11, 12, 13, 14, 15, 17, VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞/
Vi	VI	VI1, 2, 3	/*狀態類及物動詞*/
Vt	VJ	VJ1, 2, 3	/*狀態及物動詞*/
Vt	VK	VK1, 2	/*狀態句賓動詞*/
Vt	VL	VL1, 2, 3, 4	/*狀態謂賓動詞*/
Vt	V_2	V_2	/*有*/
T	DE	/*的, 之, 得, 地*/	
Vt	SHI	/*是*/	
FW	FW	/*外文標記*/	

COLONCATEGORY	/* 冒號 */
COMMACATEGORY	/* 逗號 */
DASHCATEGORY	/* 破折號 */
ETCCATEGORY	/* 刪節號 */
EXCLANATIONCATEGORY	/* 驚嘆號 */
PARENTHESISCATEGORY	/* 括弧 */
PAUSECATEGORY	/* 頓號 */
PERIODCATEGORY	/* 句號 */
QUESTIONCATEGORY	/* 問號 */
SEMICOLONCATEGORY	/* 分號 */
SPCHANGECATEGORY	/* 雙直線 */

3. 實際應用

3.1 歌詞分類

歌詞是一首歌的靈魂，它透過文字將歌曲想表達的意涵傳達給聽眾，因此我們想要就歌詞內容的分類進行研究。

3.1.1 歌詞標定

我們將歌詞分為友情、反映社會、愛情、感嘆、親情與勵志六大類別，類別分類的標準詳見 Table 2。並且由兩位專家進行 581 首歌詞的類別標定，Kappa 值為 0.854238，計算資料詳見專家標定結果統計

Table 3。

3.1.2 作法

在 581 首歌中，我們取每個類別 90%的歌曲作為 training 資料，其餘的作為測試資料。首先，利用 Log Likelihood Ratio 篩選出 500 個較有鑑別力的詞彙。由於我們有兩個以上的類別，所以是分別對每一個類別，將剩餘的五個類別的歌詞都視為一個不相關的類別來進行計算 Log Likelihood Ratio 的值，最後再將六個結果做平均成為字彙的 Log Likelihood Ratio 值，取其中前 500 個詞彙作為分類使用的詞彙。

再來，使用 Multinomial-based Naïve Bayes 的方法進行分類，每個類別的分數計算公式如下：

$$Score(c) = P(c) \prod_{1 \le k \le n_d} P(X = t_x | c)$$

Table 2. 歌詞分類標準

分類	分類標準
友情	展現好友之間的情誼，或者是祝福好朋友的歌曲。
反映社會	描述社會現象，或者社會議題的歌曲。
愛情	曖昧期、熱戀、分手、被劈腿等內容，不論內容是一個人、兩個人或兩個人以上，只要是描寫有關愛情任何階段的歌曲都屬於此分類。
感嘆	碎碎唸日常生活的無奈，或者抒發人生感嘆的歌曲。基本上此類別的歌曲是比較偏向負面情緒的。
親情	只要是跟家人相關的歌曲，不論對象是長輩、平輩或晚輩都是屬於這個分類。另外，思鄉的歌曲也屬於此分類。
勵志	激勵人正面思考，積極作為、發憤圖強的歌曲。

Table 3. 專家標定結果統計

		專家 A						
		友情	反映社會	愛情	感嘆	親情	勵志	總和
專家 B	友情	36	0	2	0	0	2	40
	反映社會	0	35	1	1	0	2	39
	愛情	3	1	331	5	1	2	343
	感嘆	0	0	4	40	2	6	52
	親情	0	0	3	1	38	2	44
	勵志	3	0	5	6	0	49	63
	總和	42	36	346	53	41	63	581

由於一首歌詞往往會涵蓋到多個類別的成分，例如療傷歌曲同時有愛情與勵志的成分，我們在分析的時候並不是武斷的將歌詞分進分數最高的類別，而是顯示這首歌的內容佔各類別的成分比例，分數轉換比例的方式為將分數除以六個分類的加總，公式如下：

$$Ratio(c) = \frac{Score(c)}{\sum_{c' \in C} Score(c')}$$

3.2 作詞人特色分析

我們選出九位作詞人（李宗盛、姚謙、林夕、瓊瑤、方文山、吳青峯、張雨生、伍佰與丁曉雯）進行差異分析。並各自挑選了每位作詞人的五十首作品作為我們的研究樣本（共計四百五十份文字檔）。而想要研究的問題是：九位作詞人在撰寫歌詞時，用字遣詞上是否有所差異？

因此，第一個步驟便是各自去造出每一位作詞人的字典，接著再依據 df 將所有出現在字典裡的詞彙作排序，得到所有作詞人使用頻率最高的十個詞彙（參照 Table 4）。

從 Table 4 我們可以得知，九位作詞人的十大最常用字是類似的，看不出明顯差異。因此，我們下一步便是去做 Feature Selection，將九位作詞人視為九個分類，使用 Chi-Square Test 去挑選出五百個最具「作詞人」鑑別力的詞彙，並依 Chi-Square 值排序，希望藉此能夠找出每個作詞人的特色慣用詞彙。排序之後，再將這五百個特徵詞彙去跟九位作詞人對應，將每位作詞人的特徵詞彙使用頻率填入下面的表格之中，即可看出各作詞人的特色用詞。

Table 4. 各作詞人使用頻率最高的十個詞彙

排序	李宗盛	df	姚謙	df	林夕	df	瓊瑤	df	方文山	df	青峯	df	張雨生	df	伍佰	df	丁曉雯	df
1	愛	32	愛	31	誰	34	心	17	愛	29	愛	26	愛	22	愛	26	愛	39
2	心	25	心	27	愛	27	夢	17	誰	28	心	24	裡	18	心	23	心	34
3	誰	25	愛情	18	像	27	裡	15	用	20	過	21	過	18	不要	18	裡	23
4	裡	21	誰	16	過	24	多少	11	風	16	像	19	心	15	看	17	不要	21
5	知道	19	世界	15	沒	22	誰	11	像	16	誰	15	生命	15	無法	17	誰	19
6	過	19	知道	15	給	18	愛	10	走	15	裡	14	走	13	裡	16	一切	16
7	像	18	裡	15	感情	16	天	9	過	15	夢	14	看	13	像	16	天	13
8	這樣	17	看	14	比	14	情	9	回憶	14	生命	13	天	11	走	14	永遠	13
9	給	16	過	14	怕	14	別	8	如	14	快樂	13	曾	11	知道	14	別	13
10	也許	12	像	13	甚麼	14	相	8	開始	14	天	12	像	11	風	14	走	12

Table 5

特徵詞彙	李宗盛(df)	姚謙(df)	林夕(df)	瓊瑤(df)	方文山(df)	青峯(df)	張雨生(df)	伍佰(df)	丁曉雯(df)	指定分類
甚麼	0	1	**14**	0	0	0	2	0	0	林夕
不要	8	7	5	2	2	4	0	18	**21**	丁曉雯
心	25	27	7	17	6	24	15	23	**34**	丁曉雯
誰	25	16	**34**	11	28	15	9	6	19	林夕
無法	8	0	2	2	6	6	2	**17**	1	伍佰

最後將分類結果，列於 Table 6（僅列出前 12 筆）。

Table 6

李宗盛		姚謙		林夕		瓊瑤		方文山		青峯		張雨生		伍佰		丁曉雯	
特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序	特徵詞彙	排序
未	36	秒	17	甚麼	1	柔情	7	搖晃	15	算	23	臭	9	無法	5	不要	2
也許	37	愛情	28	誰	4	且	24	千	18	魚	39	一天天	14	隨著	8	心	3
除了	42	微笑	92	感情	10	相思	26	字眼	21	幻想	81	徜徉	22	慢慢	20	瞭解	6
雖然	46	夠	93	沒	11	共	27	臉	31	語言	85	生命	32	全部	44	愛	19
為何	72	忽然	102	比	12	相	29	畫面	33	堅強	87	文明	41	漂浮	52	遠	68
男人	82	感想	110	情人	13	春	38	幾	34	夢境	100	談天	51	吹	67	黑夜	70
女人	86	明天	120	便	16	白雲	43	誓言	35	脈搏	113	夥	53	變成	73	永遠	90
怎麼說	95	流行	140	快樂	25	簾幽	49	家鄉	40	失眠	116	社會	57	濃霧	106	一切	129
從此	104	後來	181	令	30	珍重	55	懸崖	48	翅膀	117	東西	58	紅色	115	漂流	176
應不應該	105	流淚	213	快	45	留不住	78	颳	50	遺忘	127	白日夢	59	之中	121	滑落	177
這樣	131	海洋	225	難道	47	兒	80	風鈴	54	蝶	158	心靈	128	非常	133	上天	178
不必	136	經過	231	大概	62	相逢	91	城	56	貓咪	168	情懷	132	正在	173	值得	187

3.3 情歌用詞年代差異分析

將蒐集來的三百多首愛情歌曲，依照年代區分，以十年為一區間，分別得到 1981-1990 年 126 首、1991-2000 年 102 首、以及 2001-2010 年 113 首的歌詞作品。接著依據相同的做法，將這三類歌詞視為三個分類，各隨機取出 100 篇歌詞作為 training documents，去做 feature selection，

最後得到 500 個具有「年代」鑑別力的詞彙。最後分類後的結果列於表格 4（僅列出每個分類的前 30 筆）。

Table 7

特徵	排名	年代	特徵	排名	年代	特徵	排名	年代
輕輕	2	1981-1990	一生	13	1991-2000	幸福	1	2001-2010
似	6	1981-1990	再見	15	1991-2000	過	3	2001-2010
使	7	1981-1990	一切	18	1991-2000	愛	4	2001-2010
令	16	1981-1990	癡	34	1991-2000	陪	5	2001-2010
情意	19	1981-1990	朋友	40	1991-2000	沒	8	2001-2010
柔情	20	1981-1990	敢	50	1991-2000	多	9	2001-2010
永	30	1981-1990	所有	51	1991-2000	懂	10	2001-2010
不可	31	1981-1990	想要	54	1991-2000	看	11	2001-2010
知	32	1981-1990	潮	69	1991-2000	個人	12	2001-2010
輕	35	1981-1990	動人	73	1991-2000	近	14	2001-2010
陽光	47	1981-1990	事	75	1991-2000	起	17	2001-2010
深情	56	1981-1990	無所謂	88	1991-2000	只是	21	2001-2010
心靈	57	1981-1990	有點	97	1991-2000	比	22	2001-2010
語	70	1981-1990	感情	103	1991-2000	就是	23	2001-2010
無限	71	1981-1990	心情	105	1991-2000	下	24	2001-2010
清風	72	1981-1990	連	111	1991-2000	時間	25	2001-2010
午夜	74	1981-1990	藍天	120	1991-2000	天	26	2001-2010
離去	76	1981-1990	樣	122	1991-2000	太	27	2001-2010
風雨	78	1981-1990	緊	124	1991-2000	總會	28	2001-2010
但願	84	1981-1990	悠悠	139	1991-2000	遇見	29	2001-2010
時光	89	1981-1990	宿命	141	1991-2000	期待	33	2001-2010
分離	106	1981-1990	追求	144	1991-2000	哭	36	2001-2010
如今	108	1981-1990	看來	152	1991-2000	快樂	37	2001-2010
請	109	1981-1990	夜夜	159	1991-2000	手	38	2001-2010
悄悄	113	1981-1990	來生	160	1991-2000	假裝	39	2001-2010
歡樂	117	1981-1990	依靠	161	1991-2000	答案	41	2001-2010
飄泊	118	1981-1990	狂	164	1991-2000	給	42	2001-2010
臉龐	121	1981-1990	行李	166	1991-2000	越	43	2001-2010
溫馨	129	1981-1990	安定	169	1991-2000	放手	44	2001-2010
愛意	132	1981-1990	日	171	1991-2000	存在	45	2001-2010

3.4 歌詞產生器

由於坊間歌詞大多存在類似的句型，而只是根據不同的主題，抽換其中的詞組替換為類似的用詞，便可成為一首新歌，發現到這個現象的我們想到如果可以根據歌詞常見的句型建成 pattern，再根據這個結果把 pattern 內的用詞做替換，那麼任何一個人都可以很輕易地寫出一首歌。

於是首先我們先選出了一些暢銷作品，並從中選出讓人較為印象深刻的句子，並分析句子的組成，將句子的組成以詞性的方式分類並標記成句型 patter，以下為句型 pattern 的例子。

- *等候著[形容詞][名詞]的到來*
- *[時間]你該給的[名詞]，被你親手[副詞][動詞]*
- *你的[名詞]似乎對我訴說，[名詞]千萬不要[動詞]*
- *我願在[地點]，為你[動詞][名詞]，又怕你[副詞]遠去，讓[名詞]笑我[形容詞]*
- *我決定，離開[地點]，哪怕會有[動詞]的可能*
- *我的[名詞]一[動詞]就看到[名詞]*

而我們也發現到，大部分愛情歌曲的用詞其實也極為相近，於是考慮選擇大部分愛情歌曲會用到的詞組作為替換的詞組來源，根據我們對於「愛情歌曲年代差異分析」，選出的 500 個具有「愛情歌曲年代」鑑別力的詞組，以這 500 個詞組作為句型內用詞替換的用詞來源，會讓產生出來的作品帶有愛情歌曲的感覺。而由於是根據詞性去抽換的，句型本身會是通順的，僅有可能在邏輯上產生怪異的狀況，但是因為坊間歌詞內容本身即為天馬行空的內容，因此即使邏輯上出現怪異的結果也不會距離真實的歌詞太遠，達到了自動生成歌詞的目的，如下就是一個很有氣氛的生成結果。

存在

書架上的醫生，像深情歌聲，我一個人傷

你給我的聲音像世界迷離

保護吧我的名字

當時間席捲而來，這心田好像已被愛意深埋

我會陪著你飄進，陽光中會回來，親愛的你不要害怕

我的感情一編織就看到臉龐

整顆黑夜懸在城市，我只能夠癡癡轉身

朦朧的世界、深情的語，我在森林分離

4. 結論

本專題以愛情歌詞為中心，透過 Classification 及 Feature selection 等方法篩選出最能代表愛情的五百個詞彙，並利用收集而得的句型範本生成情歌歌詞，其結果相當具有娛樂效果。在現代的社會中，創意與創新已是成功的關鍵重要因素，往後若能根據使用者之心情來決定歌詞種類，必定能為此歌詞產生器更上一層樓。

5. 工作分配

工作項目	執行人員
資料收集	全組人員
中文斷詞	彭俊穎
程式撰寫	全組人員
網站架設	彭俊穎
文件編寫	全組人員

6. 參考文獻

[1] Julus, S., Chris, C. C. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* 85, 3 (March 2005), 257-268. http://www.physther.org/content/85/3/257.short

[2] McCallum, A., Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. http://www.bibsonomy.org/bibtex/15e99bd172bb97fc446913878ae78c233/ngrandy)

[3] Ted, D. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19, 1 (March 1993), 61-74. http://dl.acm.org/citation.cfm?id=972454