MACHINE LEARNING HW4 REPORT

B02901065 電機四 李洺曦

1. 使用sklearn中的TfidfVectorizer作為提取features之frequency的工具，將use_idf設為False，使用並且將max_features設為10，藉此提取出出現頻率最高的10個words，以下為統計結果：

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
wordpress	to	to	with	with	to	visual	with	to	with	with	with	with	web	with	with	with	with	type	to

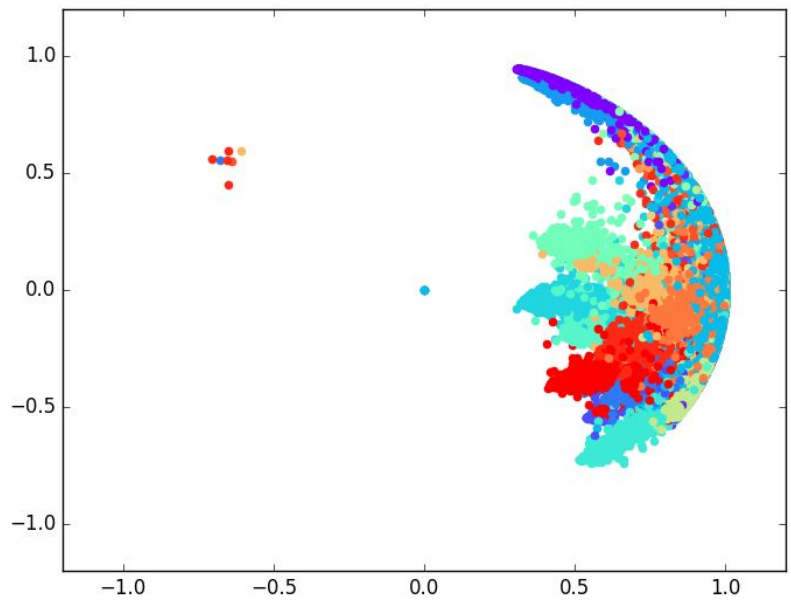
接者將irrelevant words提出，將stop_words設為"english"，使用sklearn內建的dictionary for stop words，並將max_df設為0.5，忽略frequency高於0.5的words，得到的結果如下：

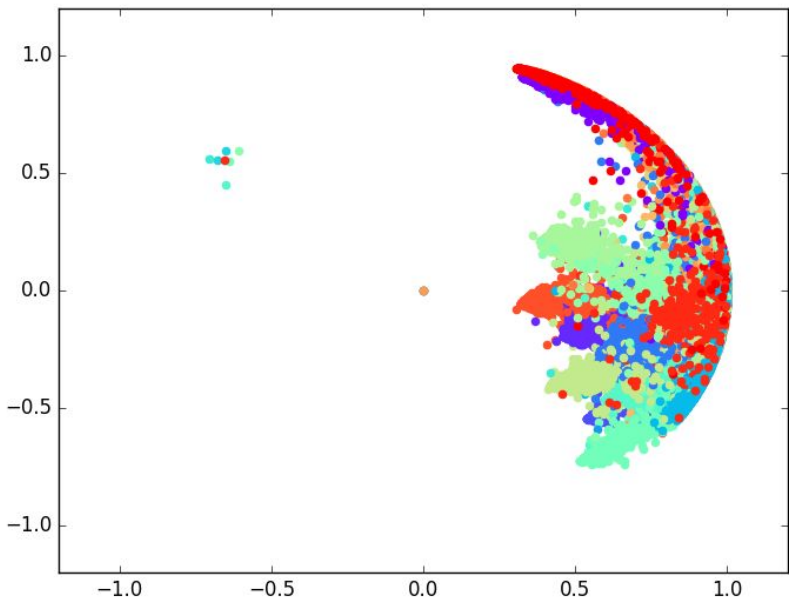
1	2	3	4	5	6	7	8	9	10
using	using	using	using	vba	vector	vs	window	using	variable

11	12	13	14	15	16	17	18	19	20
xml	using	way	web	web	windows	views	xml	using	products

可以發現到字群不再以"to", "with"等常用且無意義的英文字符為大宗，而是多出了一些新的字，但值得注意的是第1群，在沒有使用max_df以及stop_words時頻率最高的字恰好就是label，但是使用後就被取代掉了，推斷是因為太常出現，所以被max_df過濾掉。而using的出現，可以使用nltk.snowball之類的套件將英文的變化形轉回原形，並再調低max_df以過濾掉這些重複出現在各cluster的字。

2.





左圖為自行分20群，右圖為真實標籤，可以發現在kmeans在分群上分的很平均，沒有橫跨區域的問題，而真實label有覆蓋以及橫跨的現象產生，推測是在抽feature時沒有處理好，導致維度不足以描述群與群的空間關係。

3. 這裡比較4種feature extraction methods，分別是Bag of Words with stop_words、Bag of Words without stop_words、Tfidf with stop_words、Tfidf without stop_words（使用tri-gram, max_df=0.95, min_df=2）

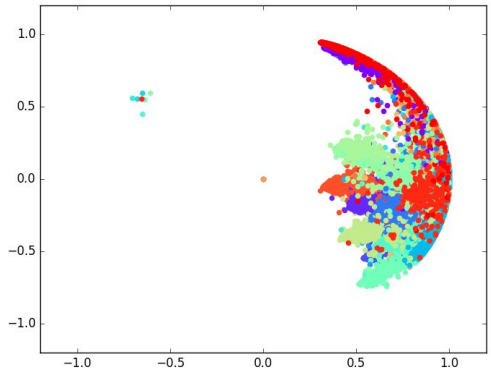
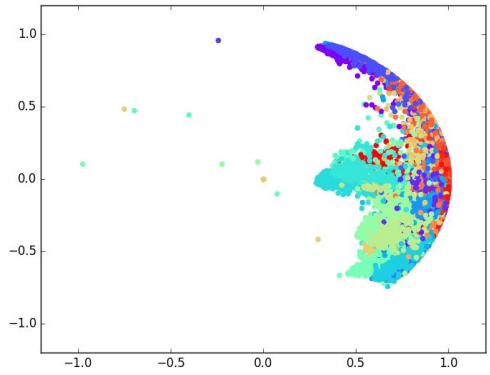
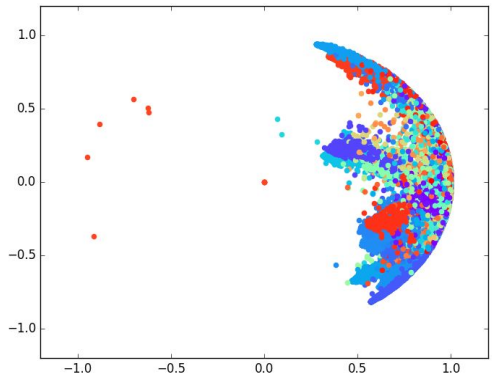
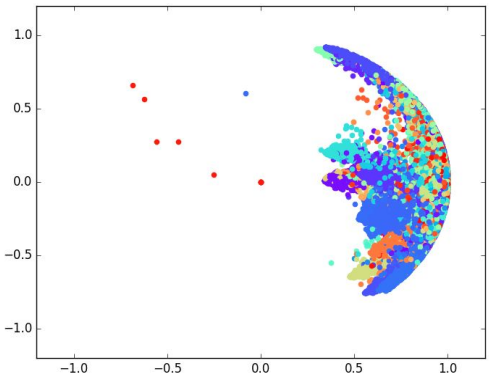
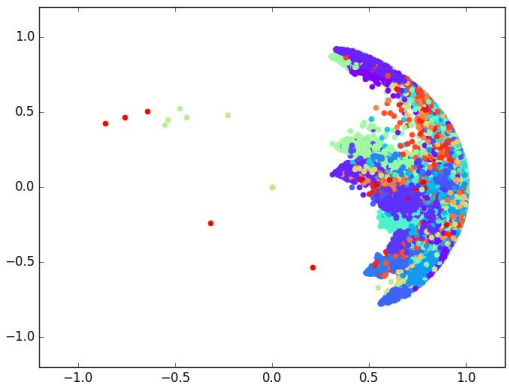
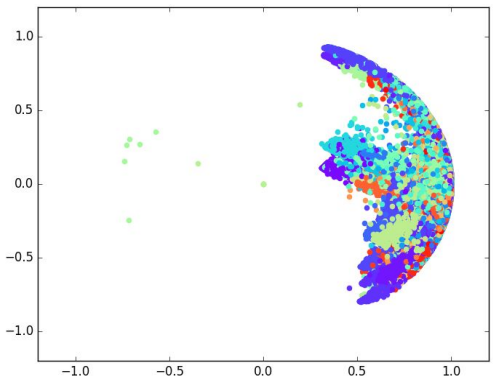
	features count	public set	private set
BoW with SW	1776334	0.15201	0.14846
BoW without SW	1387145	0.8332	0.83209
Tfidf with SW	1776334	0.59977	0.59237
Tfidf without SW	206271	0.8793	0.87902

可以發現，有去除stop words與沒去除stop words的差距很大，vector的數量也有差距，而使用Tfidf會比單純使用Bag of Words還要來的有意義。

4. 下表為各clusrer number之分數，可以發現當分成80份與120份時分數最高

clusters count	public set	private set
20	0.54702	0.5464
30	0.80768	0.80296
40	0.84894	0.8478
50	0.87376	0.87245
60	0.88086	0.88086
70	0.8888	0.88777
80	0.89104	0.8917
90	0.88669	0.88459
100	0.88882	0.88935
110	0.88793	0.88676
120	0.89099	0.89172
130	0.88127	0.88073
140	0.88548	0.88458
150	0.88291	0.88285
160	0.87588	0.87495
170	0.87614	0.8768
180	0.87873	0.87988
190	0.87853	0.87864
200	0.87701	0.87867

下圖為cluster number取真實label、40、80、120、160、200時的二維分群圖演變

可以觀察到由於cluster數目的增加，可以將真實label中跨過另一個cluster的cluster分為兩個cluster平均的切割，以達到叫為準確的預測。