機器學習(HW4)　姓名：袁培傑　學號：B03901134

1. **Analyze the most common words in the clusters. Use TF-IDF to remove irrelevant words such as “the”.**

我的 Model 是如此運作的，首先利用 nltk 將 title、doc 去掉標點符號，大寫轉換為小寫，取出每個單字的語幹，再餵給 TF-IDF，參數設置 max_df=0.4, min_df=2，然後會產出 20000×28289 的 vector，再丟進 SVD(LSA)降至 20 維，再以這些 vector 去做 Kmeans，得到 20 組 cluster，去計算各個 cluster 出現的單字數，左圖為沒有去掉 stopwords 的數目，而右圖是已經去掉的情況。可以發現各組幾乎可以將 20 個真正的 tag 偵測出來。

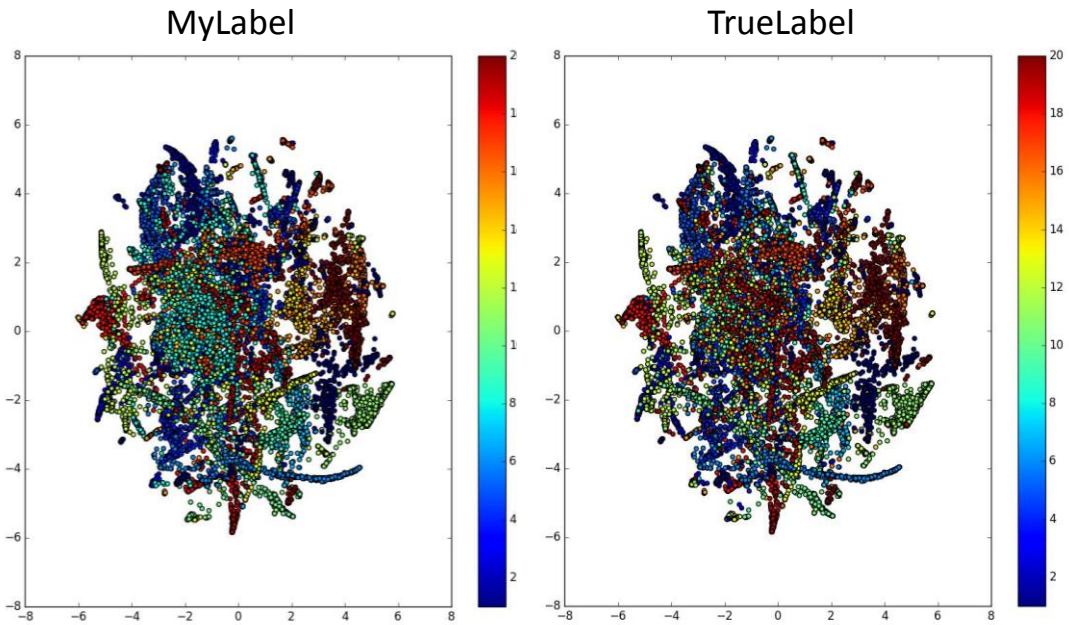
Cluster 0 : visual: 719 studio: 692 in: 326	Cluster 0 : linq: 856 queri: 198 sql: 166
Cluster 1 : sharepoint: 741 a: 343 to: 278	Cluster 1 : drupal: 855 view: 171 node: 147
Cluster 2 : to: 869 a: 737 in: 533	Cluster 2 : excel: 877 vba: 141 cell: 130
Cluster 3 : spring: 832 to: 239 in: 219	Cluster 3 : ajax: 743 jqueri: 113 call: 73
Cluster 4 : bash: 676 a: 468 in: 410	Cluster 4 : cocoa: 330 subvers: 268 object: 154
Cluster 5 : drupal: 854 in: 315 to: 308	Cluster 5 : wordpress: 874 post: 244 page: 171
Cluster 6 : ajax: 749 to: 203 in: 146	Cluster 6 : sharepoint: 745 web: 117 custom: 105
Cluster 7 : apache: 609 to: 331 rewrite: 201	Cluster 7 : spring: 843 bean: 113 hibern: 82
Cluster 8 : magento: 874 in: 333 to: 277	Cluster 8 : apach: 604 rewrit: 212 mod: 190
Cluster 9 : excel: 876 in: 395 to: 394	Cluster 9 : magento: 877 product: 237 custom: 108
Cluster 10 : scala: 804 in: 350 a: 241	Cluster 10 : hibern: 839 map: 136 queri: 83
Cluster 11 : svn: 622 to: 397 a: 311	Cluster 11 : haskel: 734 function: 194 thi: 36
Cluster 12 : linq: 854 to: 439 a: 247	Cluster 12 : matlab: 829 array: 92 function: 88
Cluster 13 : matlab: 828 in: 462 to: 281	Cluster 13 : visual: 723 studio: 695 2008: 141
Cluster 14 : wordpress: 871 to: 323 in: 265	Cluster 14 : bash: 678 script: 299 command: 132
Cluster 15 : haskell: 729 in: 389 a: 252	Cluster 15 : qt: 607 window: 118 widget: 67
Cluster 16 : hibernate: 859 to: 300 in: 218	Cluster 16 : svn: 618 repositori: 91 commit: 61
Cluster 17 : oracle: 761 to: 335 in: 281	Cluster 17 : scala: 806 java: 85 class: 83
Cluster 18 : mac: 481 os: 344 x: 302	Cluster 18 : oracl: 770 sql: 197 tabl: 98
Cluster 19 : qt: 608 in: 249 to: 227	Cluster 19 : mac: 490 os: 341 x: 301

2. **Visualize the data by projecting onto 2-D space. Plot the results and color the data points using your cluster predictions. Comment on your plot. Now plot the results and color the data points using the true labels. Comment on this plot.**

LSA 將 20000 筆 data 降至 20 維後，透過 TSNE 把這些 data 投影到 2D 平面上，並且利用 true label 的答案，將 Kmeans 分出來各 cluster 的編號配對後，形成左右兩張圖，左圖為自己 label 的結果，右圖為真正的 Label。

可以發現，會形成幾條線以及幾群，代表那組 cluster 算是有分出來的情況；而正中心青藍色的 label 明顯與實際的 label 有落差，此 tag 是 cocoa、或者是有時候 20 個 cluster 分不出明顯的 tag 而將一些不相關的 data 都丟入此 cluster，導致利用 20 個 cluster 時正確率低落。透過 title 去觀察發現 cocoa 常常

與其他字詞混在一起，像是 mac 等等，導致 20 個 cluster 不容易去分析到這個 tag。



3. Compare different feature extraction methods.

我利用 BoW、TF-IDF 做不同的 feature extraction，都同樣丟入 LSA 降到 20 維，並且兩者會分別去觀察去除 Stopwords 的分數差別，在輸出答案時是利用 cos similarity > 0.9 設為 1 的方式去實現，沒有去除 Stopwords 的 feature 為 28415，而去除後為 28289。結果如下，可以發現同樣沒有去除 Stopwords 的情況下，TF-IDF 表現遠勝 BoW，原因可能是因為 BoW 接者做 LSA 使用 cos simlarity，如此的方式會導致 LSA 做完沒有辦法將 20000 筆 data 分很開，沒有得到想要分群的效果。在去除 Stopwords 後，兩者成績有顯著提升。

	BoW		TF-IDF	
Stopwords	Public	Private	Public	Private
With	0.90088	0.90044	0.91632	0.91604
Without	0.15724	0.15063	0.65072	0.64545

4. Try different cluster numbers and compare them. You can compare the scores and also visualize the data.

比較第二題的圖(上圖)以及下圖，前者是分成 20 個 cluster，後者是分成 100 個 cluster，為了避免 100 個 cluster 顏色不清，我將 100 組出現最多次的 tag 與

true label 進行配對，減少至 20 個 tag 進行分析。兩圖互相比對可以發現兩個特點：首先，觀察中間區塊，後者的圖較鬆散，而前者的圖較緊密，且後者與 true label 比較沒有太大的誤差；第二，後者有較多的線段產生，代表 100 個 cluster 會比 20 個 cluster 分的較開，實際上分數從 0.65 提升 0.80，我猜想可能與分數的計算方式有關，我們要盡可能猜對 1 的存在，減少猜成 1 的可能，所以分成 100 組可以提升猜對 1 的可能。

然後我試著不用 Kmeans 進行分群，而只透過 cos similarity > 0.9 進行兩兩匹配，成績直接躍升 0.9，因為此次作業目標是檢測兩者是否同群，而不是綜觀全體分成 20 組，每組正確率越高分數就越高，因為目的不同的關係，利用 Kmeans 反而會產生錯誤比較多的答案。