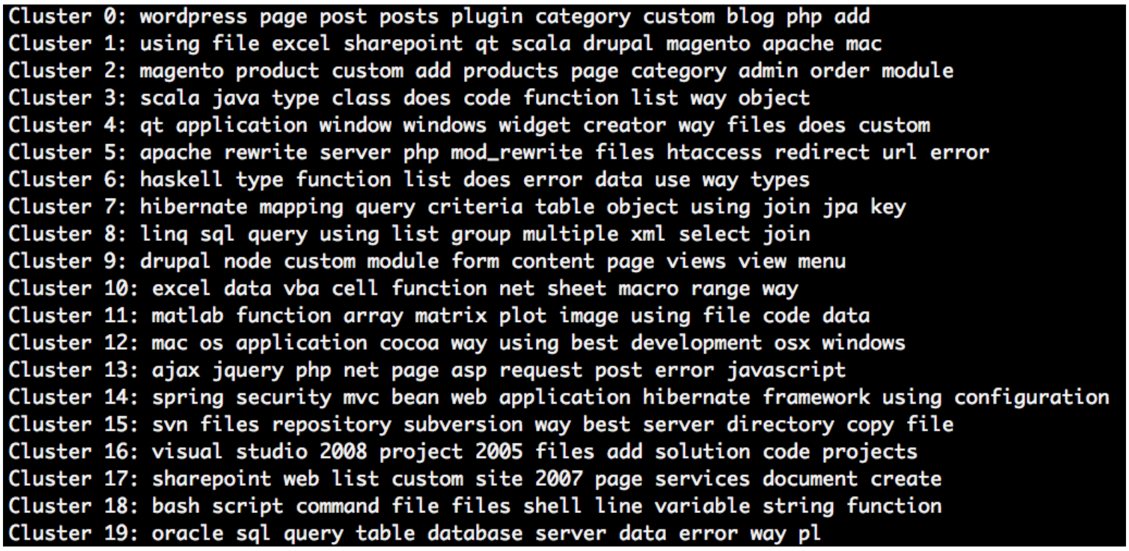
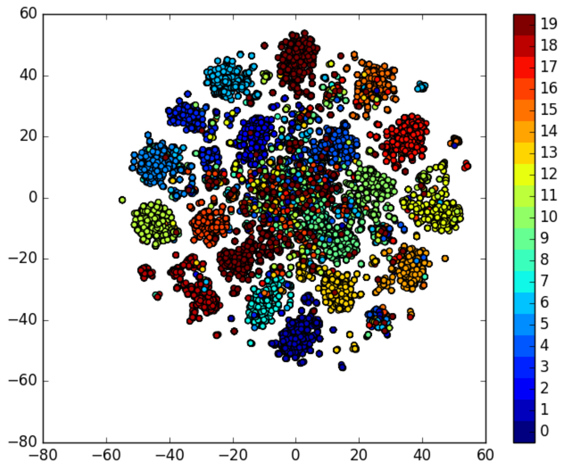
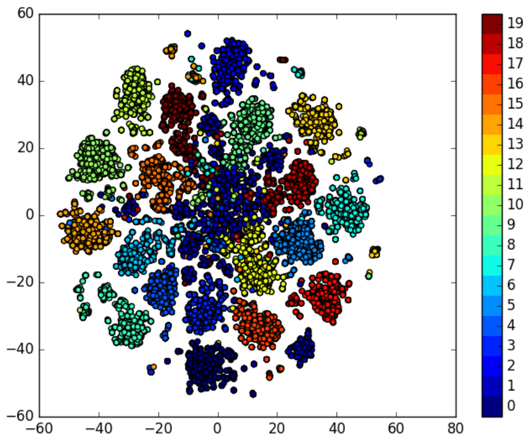
# ML HW4 REPORT b02902041 徐朝駿

1. Analyze the most common words in the clusters. Use TF-IDF to remove irrelevant words such as "the". (1%)



```
Cluster 0: wordpress page post posts plugin category custom blog php add
Cluster 1: using file excel sharepoint qt scala drupal magento apache mac
Cluster 2: magento product custom add products page category admin order module
Cluster 3: scala java type class does code function list way object
Cluster 4: qt application window windows widget creator way files does custom
Cluster 5: apache rewrite server php mod_rewrite files htaccess redirect url error
Cluster 6: haskell type function list does error data use way types
Cluster 7: hibernate mapping query criteria table object using join jpa key
Cluster 8: linq sql query using list group multiple xml select join
Cluster 9: drupal node custom module form content page views view menu
Cluster 10: excel data vba cell function net sheet macro range way
Cluster 11: matlab function array matrix plot image using file code data
Cluster 12: mac os application cocoa way using best development osx windows
Cluster 13: ajax jquery php net page asp request post error javascript
Cluster 14: spring security mvc bean web application hibernate framework using configuration
Cluster 15: svn files repository subversion way best server directory copy file
Cluster 16: visual studio 2008 project 2005 files add solution code projects
Cluster 17: sharepoint web list custom site 2007 page services document create
Cluster 18: bash script command file files shell line variable string function
Cluster 19: oracle sql query table database server data error way pl
```

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. (1%)

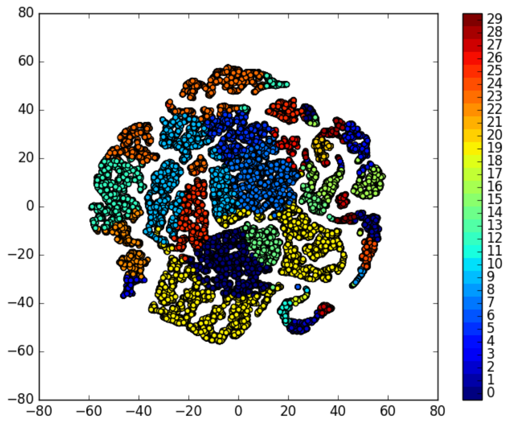
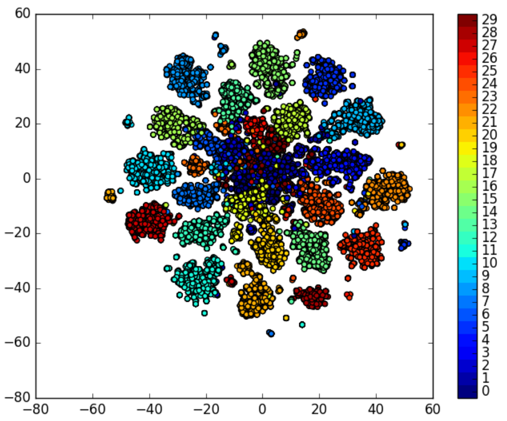
| Labeled answer | Kmeans clustering with tfidf & lsa |
|---|---|
|  |  |

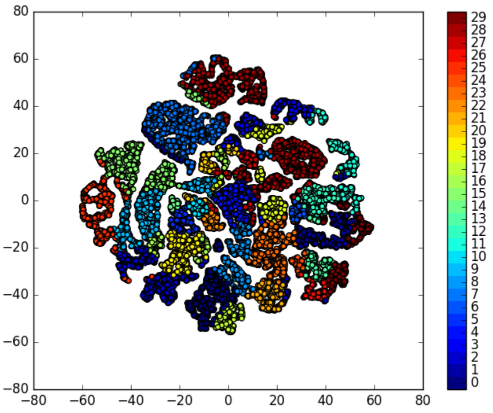
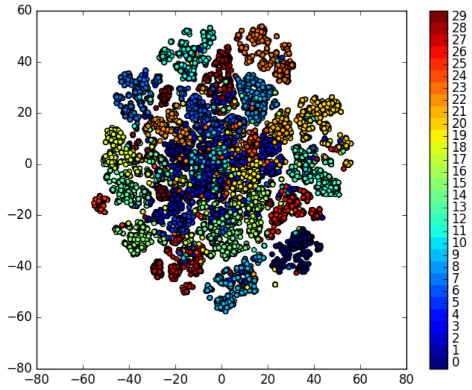
右圖是使用 tfidf 跟 lsa 抽 feature 再透過 kmeans 分類去做 predict。可以看出來除了中間有混雜的部份，四周的分類算是有效果的。不過再使用過 labeled data 去做分堆時，雖然可以看到分群的聚落分布滿相似的，但是其中有兩三群是由二至三種顏色混雜而成，這應該就是造成我的這個分類只有 0.65 正確率的原因

3. Compare different feature extraction methods. (2%)

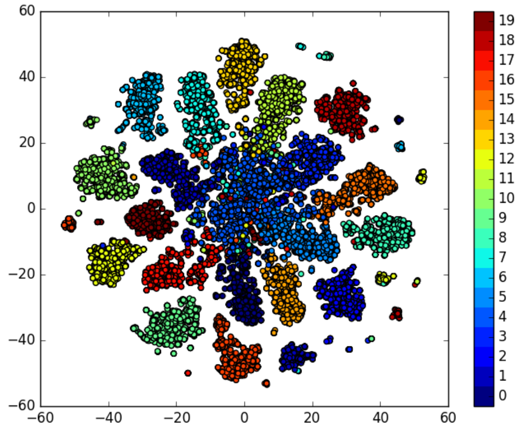
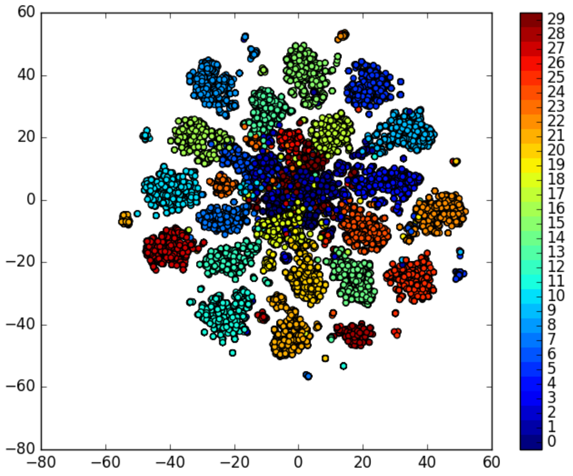
pca 的成績遠遠低於 lsa，原因可以從圖形中發現，pca 的分群不像 lsa 一樣的乾脆，可以明顯看出群落來。比較像是從一群相連的點之中，硬是劃分出 30 個分群，導致正確率很低，只比 random 好一些。另外可以看出有過濾掉 stop words 也對成績有幫助，從 lsa 圖看得出來，過濾 stop words 的分群更為清楚，不會有太多的孤島存在。

| Feature extraction | Accuracy |
|---|---|
| tfidf+pca | 7% |
| tfidf+stop words+pca | 14% |
| tfidf+lsa | 47% |
| tfidf+stop words+lsa | 77% |

| PCA with stop words | LSA with stop words |
|---|---|
|  |  |

| PCA without stop words | LSA without stop words |
|---|---|
|  |  |

4. Try different cluster numbers and compare them. You can compare the scores and also visualize the data. (1%)

| 20 Clusters | 30 Cluster |
|---|---|
|  |  |
| Accuracy 65% | Accuracy 77% |

從左圖中可以發現，明明有一些是不同群的分類，卻因為分類數量不夠多的緣故，導致會將不一樣的分類當做同一個，所以我在稍微數過重複顏色的群後，增加分群至 30 類，就讓結果有不小的上升。