# Machine Learning: HW4 Report

December 9, 2016

*Professor Hung-Yi Lee*

電機三 B03901018 楊程皓

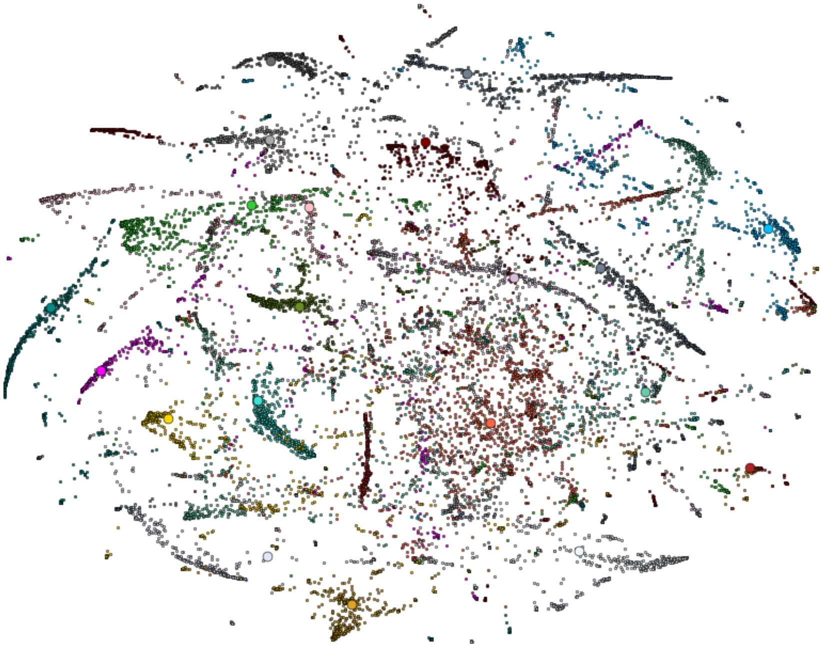
## 1. Analyze common words

使用 nltk stopwords 將很常出現且較無意義的字濾掉，再使用 TF-IDF 濾掉 stopwords 漏掉的其他多餘的字。

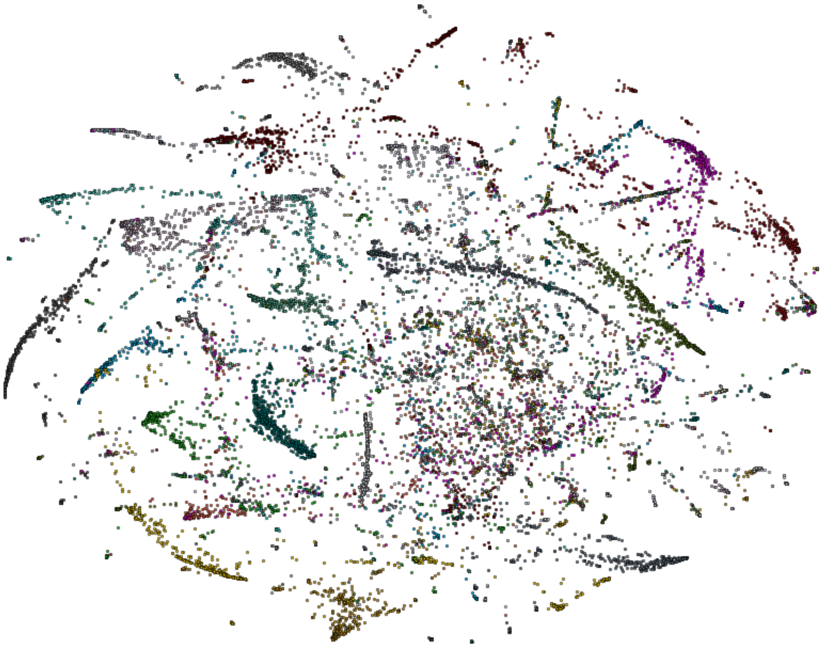
stopwords: i me my myself we our ours ourselves you your yours yourself yourselves he him his himself she her hers herself it its itself they them their theirs themselves what which who whom this that these those am is are was were be been being have has had having do does did doing a an the and but if or because as until while of at by for with about against between into through during before after above below to from up down in out on off over under again further then once here there when where why how all any both each few more most other some such no nor not only own same so than too very s t can will just don should now d ll m o re ve y ain aren couldn didn doesn hadn hasn haven isn ma mightn mustn needn shan shouldn wasn weren won wouldn

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

## 2. Visualize onto 2-D space



上圖為 cluster number = 20 的結果所作的 2D 圖，可以看出除了中間偏右側區域的點比較散亂以外，大體而言算是分的比較清楚。



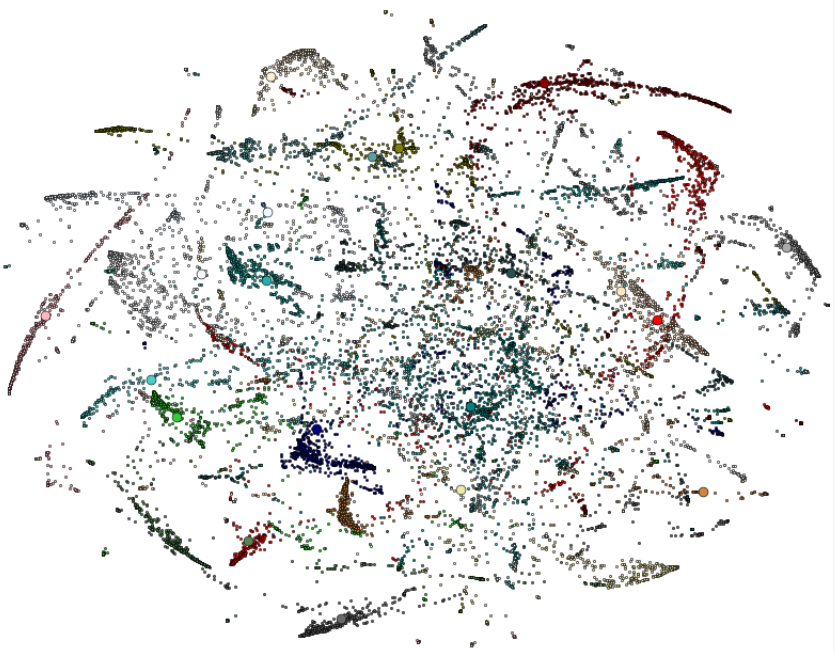
上圖為以正確的 label 作圖的結果，與前一張圖結果差不多。兩圖的結果說明中間偏右在本次的 cluster 中分的不清楚，其餘部份結果還算不錯。

## 3. Compare different feature extraction methods

一開始使用的是 bag of words + TF-IDF, 並調整 cluster number 與給定不同的 stopwords, score 約在 0.2 0.36 間。
後來加上 LSA，score 即上升到約 0.8 左右，調整各參數，包括 LSA所需的 n_components，可以讓 score 上升到 0.85 左右。
由此可見，語意上的分析在 title 識別上也有很大的影響。

## 4. Different cluster numbers

嘗試過四種不同 cluster number: 20, 70, 85, 100.
public score 分別約為: 0.636, 0.856, 0.850, 0.801.
故依照上述結果，作 cluster number = 70 的 2D 圖如下:



需要這麼多 clusters 代表每個 cluster 有許多關鍵字，而每個 title 並不一定會涵蓋到所有，只有使用更多其他 data, feature extractions 得知關鍵字間的關係，才能辨別這些 clusters 的相同, 相異。