Machine Learning HW4 Report

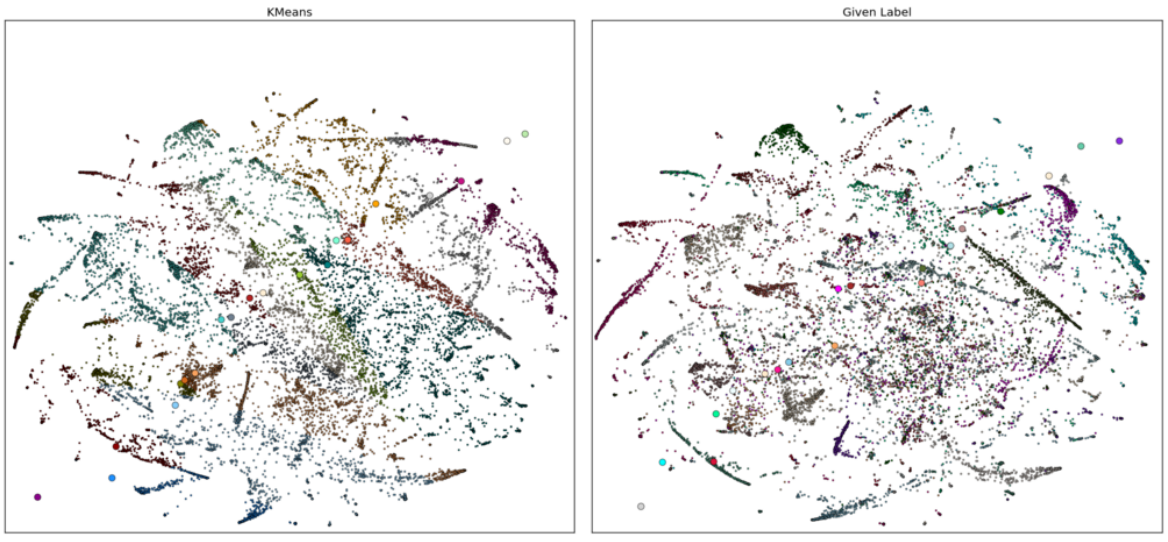
B03901034 吳泓霖

Common Words

When reading in the titles and docs, I lower cased all words, removed all punctuations, removed all stop words (imported from `nltk.corpus` package) before doing further operations.
The following are the common words for each clusters:

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

Visualise onto 2D Graph



The two graphs shows insignificant differences between the predicted clustering and the accurate answer.

Different Feature Extraction Methods

I tried the difference between whether use LSA, and to no suprise, using LSA is significatly better than the case not using it.
Here are the private score of the cases with LSA and without LSA, clustering into 70 groups:

Without LSA	With LSA
0.34113	0.85208

Different Cluster Numbers

I experimented with cluster number of 20, 50, 70, 100, and compare them with the private scores.
In general, more clusters get a better score, but too much (e.g. 100) will eventually reduce the performance.

20	50	70	100
0.63479	0.83809	0.85208	0.84295