Machine Learning
Unsupervised Clustering & Dimensionality Reduction

B02901031 陳緯哲

- **Analyze the most common words**

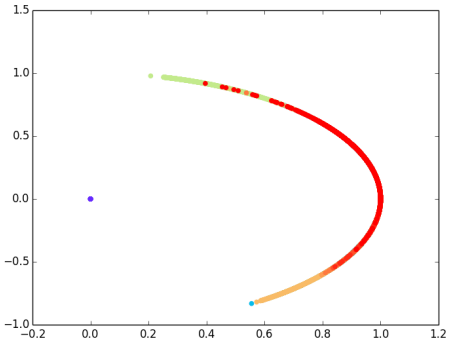
  My cluster : magento, matlab, qt, hibernate, drupal, scala, sharepoint, oracle, excel, apache, ajax, linq, visual, bash, spring, svn, use, wordpress, file, haskell

  True tags：wordpress, oracle, svn, apache, excel, matlab, visual-studio, cocoa, osx, bash, spring, hibernate, scala, sharepoint, ajax, qt, drupal, linq, haskell, magento

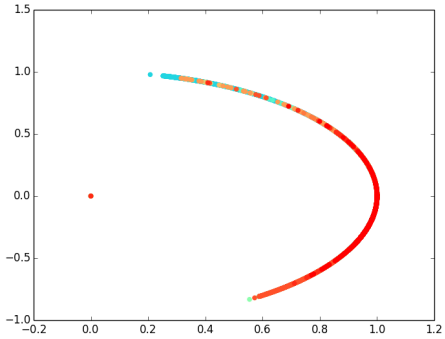
  Discussion：The most frequent term in the data are mostly same with the tags of the title. While some terms are not quite correct, such as “use”, “file”. These words are frequently appear in the title. However, they are not unique as a tags. If we tune the parameters to eliminate more frequent terms in the title, maybe the result will be more correct.

- **Visualize the data**

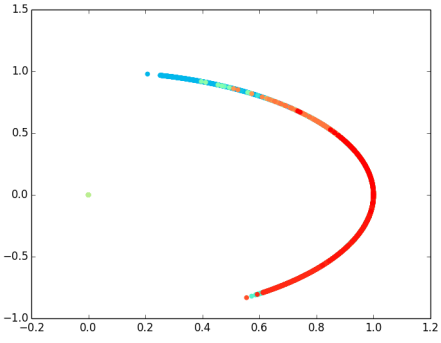
  My cluster (20 cluster)



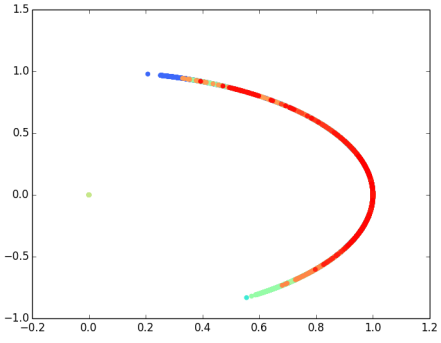
  True cluster



  40 cluster



  120 cluster



Method : I use TruncatedSVD to reduce the dimension to 2D so that I can plot the data distribution. While I use the label that are cluster by using 20 dimensions.

Commend : The similarity between two distributions are high. While some data are labeled wrong in my cluster.

- **Compare different methods**

Normalization of TF/IDF: The original tf/idf are in different scale that are hard to cluster them. By normalization, I improve the accuracy from 0.3129 to 0.6020.

Remove Punctuation: Some title has punctuations that are hard for us to cluster. I use the nltk toolkit to remove them.

TF/IDF experiment on punctuation:

with punctuation	Without punctuation
0.6102	0.6411

Remove stop word : Remove some words appears in high frequency but is useless for us to classify the data, such as “the”,” to” and “also”.

Stem : There are some morphological affixes in English, we may want to remove this affixes to enhance our performance.

Bag-of-Word vs TF/IDF without removing stop word and stem

	Cluster=20	Cluster=100
BoW	0.1041	0.1384
TF/IDF	0.3669	0.5011

Bag-of-Word vs TF/IDF with removing stop word and stem

	Cluster=20	Cluster=100
BoW	0.5730	0.8021
TF/IDF	0.6523	0.8210

- **Different cluster**

cluster	20	30	100	120
Accuracy	0.6371	0.7591	0.8097	0.7942

F measure:

$$\mathrm{P} = \frac{TP}{TP+FP} \quad \mathrm{R} = \frac{TP}{TP+FN} \quad F_\beta = \frac{(1+\beta)^2 \times P \times R}{(\beta^2 \times P)+R}$$

Increasing the cluster number can decrease the FP, making the accuracy higher. However, the if we tune the cluster number too high(cluster=120), the FN will go higher too. And the accuracy goes down.