Machine Learning HW4 Report

b02901122 電機四 劉致廷

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

   A: 對於字串的前處理，我移除了標點符號，移除了 stop words，接著利用了 TF-IDF 去 vectorize 每一行句子，經過 lsa 降低維度，丟入 k-means 分成 20 群後，找出了每一群中最常出現的字：**0. apache 1. oracle 2. svn 3.cocoa 4.excel 5.scala 6.mac 7.wordpress 8.bash 9.matlab 10.visual 11.hibernate 12.linq 13.ajax 14.spring 15.magento 16.drupal 17.sharepoint 18.qt 19.haskell**

   如果拿來跟我們的正確 tags 比較

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

   如果把’visual-studio’對應’visual’，’osx’對應’mac’，可以發現是完全對應的!代表每一群中的高頻字，其實是有 match 到我們的正確 tag，也代表分群其實有大致區分出來的。

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.**

   A: 我利用 LSA 降維成 2-D，並且用 matplotlib plot 出點圖，可以發現分佈為一個弧形，第一張是 k-means 分群後的標示圖，而第二張則是真實 label data 的圖，在第二張圖片中可以發現，其實很多我們的分群並不太正確，在降成二維後，有許多不同群的 data 被 project 到附近了！

   figure 1: plot by our cluster , and different color or shape represent different cluster.

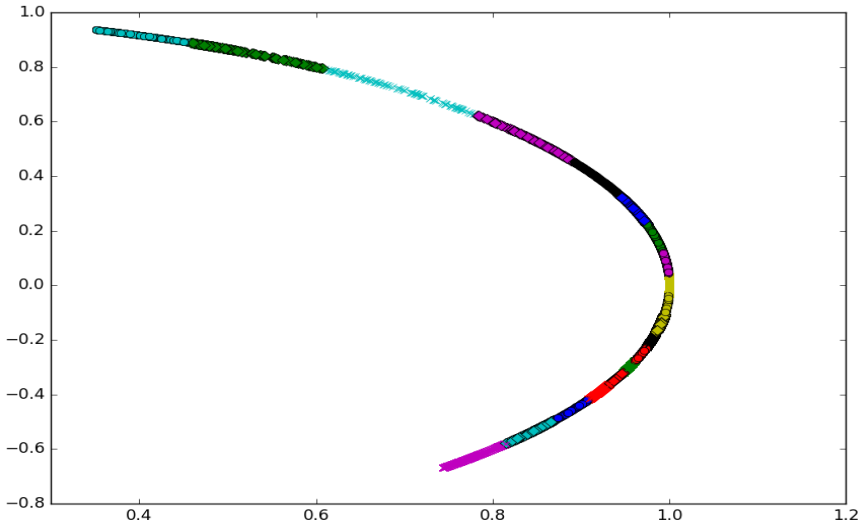
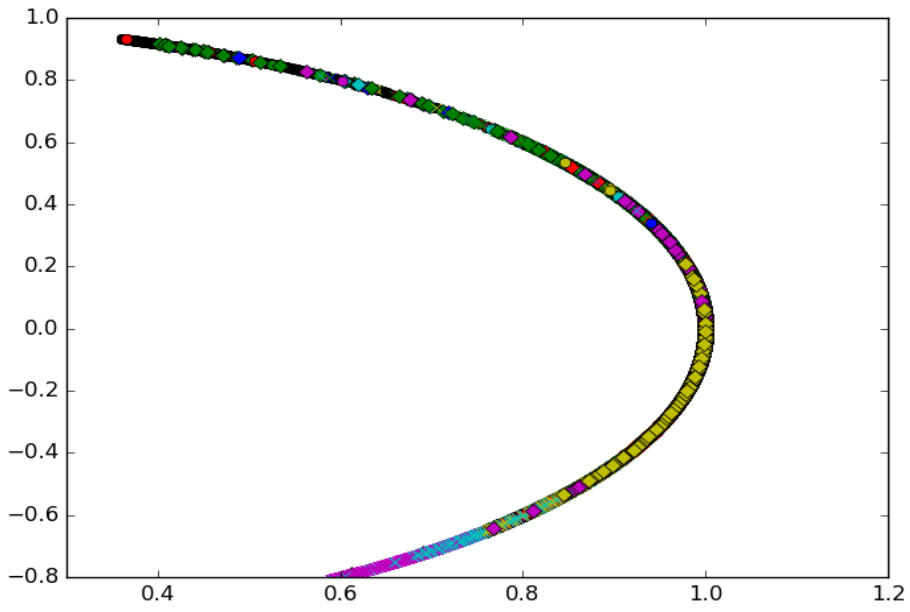


figure 2:　plot by real label , and different color or shape represent different cluster.



3. **Compare different feature extraction methods.**

   文字的前處理都統一移除掉標點符號,因為沒有固定 seed,所以可能每次出來的答案會有一點點差距。

   First method: TF(BoW,no stop words)+LSA (n_component=20)→k-means(n_cluster=20)

   A: performance: 0.10614　（異常的爛！）

   Second method: TF(BoW,stop words)+LSA (n_component=20)→k-means(n_cluster=20)

   A: performance: 0.57182

   Third method: TF-IDF(max_df=0.5,min_df=2,no stop words)+LSA(same)→k-means(same)

   A: performance: 0.32845

   Fourth method: TF-IDF(max_df=0.5,min_df=2,stop words)+LSA(same)→k-means(same)

   A: performance: 0.5963

   結論：把 IDF 加進去的確會使 performance 上升，但把 stopwords 移除會使 performance 上升很多，因為可以不用計算像是常出現的'what','who','some','such','of'.....等等。

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

   A: 文字前處理為移除標點符號、stop word，所有需要 seed 的地方都固定。

   結果如下表

Cluster	20	60	80	100	120
Performance	**0.58382**	**0.82256**	**0.83463**	**0.84173**	**0.83768**

cluster 大約在 100 的時候 performance 最高，我認為增高 cluster 數目會使 performance 上升的原因是因為我們判斷的標準是判斷倆倆是否為同一群，而其實真實 data 中大部分兩兩都不屬於同一群，因此分得越多群只是會讓判斷出來的答案有虛擬上的增強！