Machine Learning HW#4 Report

Unsupervised-learning 電機四 b02901126 張大方

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

   Tags:

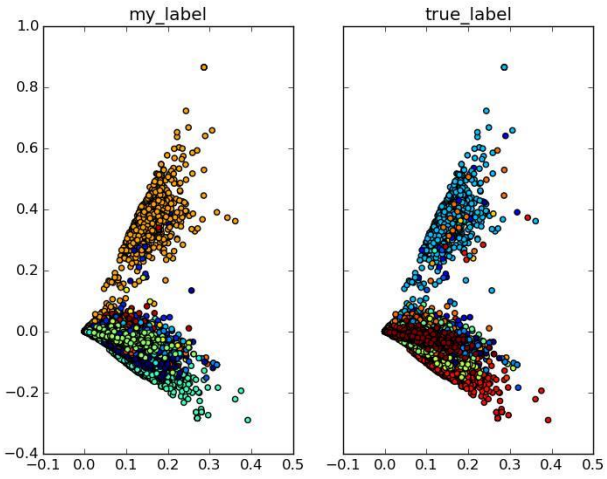
Wordpress	Oracle	Svn	Apache	Excel
Matlab	Visual-studio	Cocoa	Osx	Bash
Spring	Hibernate	Scala	Sharepoint	Ajax
Qt	Drupal	Linq	Haskell	magento

   My results:

Wordpress	Oracle	Svn	Apache	Excel
Matlab	Visual	Using	mac	bash
spring	Hibernate	Scala	Sharepoint	Ajax
Qt	Drupal	Linq	Haskell	Magento

   Each title has a vector to keep the score of key words. After summing all vectors in the same cluster, the most common word of that cluster is then chosen by the highest score in the vector.

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.



   I use TruncatedSVD to reduce the dimension down to two so that it is easier to plot. Despite that some information are lost, we can still learn from the figure that the shape of each cluster is similar between my_label and true_label, which means that my prediction model has a pretty good performance.

3. Compare different feature extraction methods.

(1). Stopword

No stopwords, cluster = 100

BoW	0.16069
Tfidf	0.44720

With stopwords, cluster = 100

BoW	0.84575
Tfidf	0.85040

(2). Latent Semantic Analysis (with stopword)

No LSA

BoW	0.47428
Tfidf	0.35997

With LSA

BoW	0.84575
Tfidf	0.85040

Both stopwords and LSA are indispensable to this task because the influence of both are significant. With stopwords enabled, words like ‘the’ are removed from the data. LSA reduces the dimension of word vectors and find relationship between words.

Comparison of the performance between BoW and Tfidf is ambiguous, but Tfidf reaches the highest score.

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

Clusters	20	50	100	150
Score	0.63108	0.83316	0.85040	0.81822

Because most of the given data is not in the same cluster, I can intuitively say that more clusters separates the titles more, enhancing the performance of prediction. In a more precise way to explain this, increasing the number of clusters can decrease FP, and then increase the accuracy. However, FN also increases with number of clusters, which leads to the loss of accuracy at the same time.