Machine Learning

Homework #4 - Unsupervised Learning

Clustering & Dimensionality Reduction

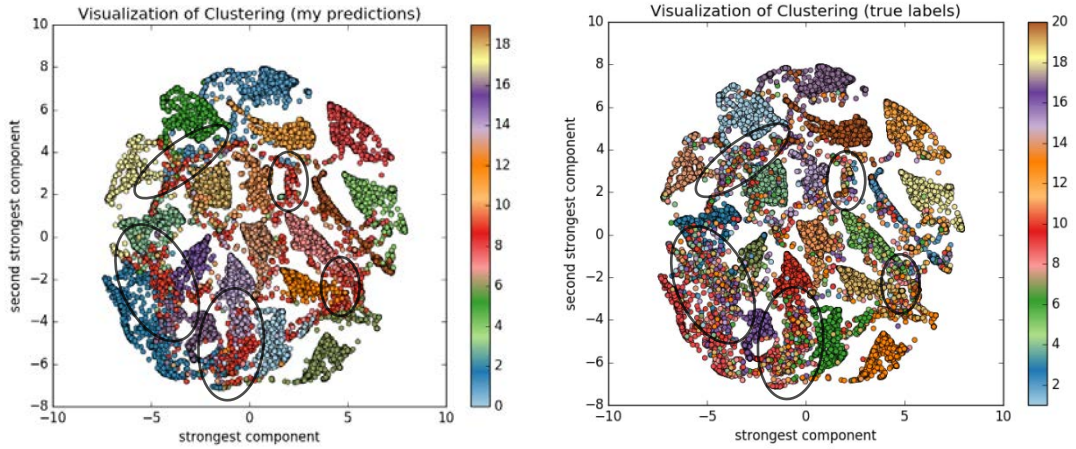
B02901080 電機四 董皓文

Problem 1

Table 1. Most common words before removing stop-words

an	and	can	do	excel
for	from	hibernate	how	in
is	magento	of	on	the
to	using	what	with	wordpress

Problem 2



(a) my predictions (b) true labels

Fig. 1 Visualizations of K-Means clustering using T-SNE mehod
(Note that the color of a tag in (a) may not be the same as in (b).)

Discussion

Since the K-Means clustering and T-SNE are computed based on the **same** LSA features we extract from all the article titles. **Thus samples in the same cluster stay closer to one another in Fig. 1(a)**, in which the color represents my prediction of the label of each sample, compared to Fig. 1(b), in which the color represents the true label of each sample.

Compared the two figures, most of my prediction errors occur on the samples in orange, as I circle out in the figures. **This indicates that the features I use as input to the K-Means clustering is not sufficient to distinguish that specific tag.** And it might be the reason why I got only an **86% accuracy** at the end.

Problem 3

Feature extraction method	F-score
BoW	0.0694
BoW (stop words)	0.1424
BoW (stop words + stemmer)	0.1749
TF-IDF (stop words)	0.1600
TF-IDF (stop words + stemmer)	0.2430
LSA on TF-IDF (stop words + stemmer)	0.4943

clustering method: K-Means clustering with 20 clusters
stop-words list: NLTK stop word list and my observations
stemmer: Porter Stemmer from NLTK

Discussion

Before removing stop-words, BoW only get an accuracy of 7%, which is slightly beyond random guess (5%). By removing stop-words and using Porter stemmer, the accuracy increases to 18%. (A stemmer remove the suffix of a word, leaving the word stem only.)

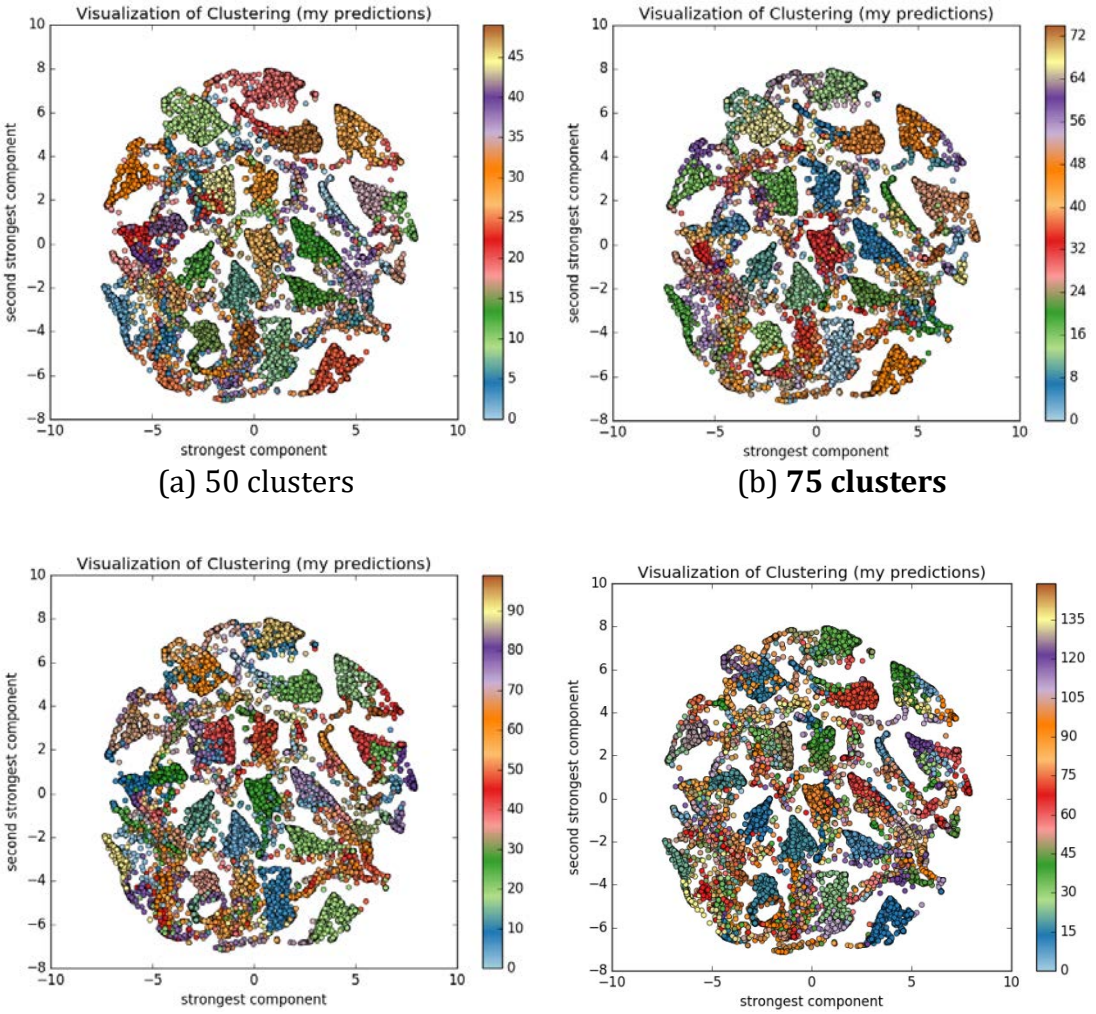
By using TF-IDF to put different weights on different words, the accuracy increases to 24%. However, the dimensions of a TF-IDF vector is too high if we want a better accuracy via performing K-Means cluster. Thus by using LSA to reduce the dimension of TF-IDF vectors and use the computed LSA feature vectors as the input of K-Means clustering, a 49% accuracy can be achieved.

Note that in this part, the K-Means clustering is computed in 20 clussters, thus having a lower accuracy in F-score.

Problem 4

# of clusters	20	30	60	**75**	80	100
F-score	0.4943	0.7960	0.8529	**0.8647**	0.8601	0.8598

clustering method: K-Means clustering
stop-words list: NLTK stop word list and my observations
stemmer: Porter Stemmer from NLTK



(a) 50 clusters (b) **75 clusters**

(c) 100 clusters (d) 150 clusters

Fig. 2 Visualizations of K-Means clustering with my predictions using T-SNE. (Note that the color of a tag may differ in different figures.)

Discussion

The highest F-score is achieved using 75 clusters, in which some small regional clusters are detected but not too fragmentized. Lower cluster numbers cannot detect some small regional clusters, thus combine them to a wrong cluster. However, a too-high cluster numbers may generate too many small clusters, and loss the general picture in a whole. The clusters might be too fragmentized, and it might result in a reduced score.