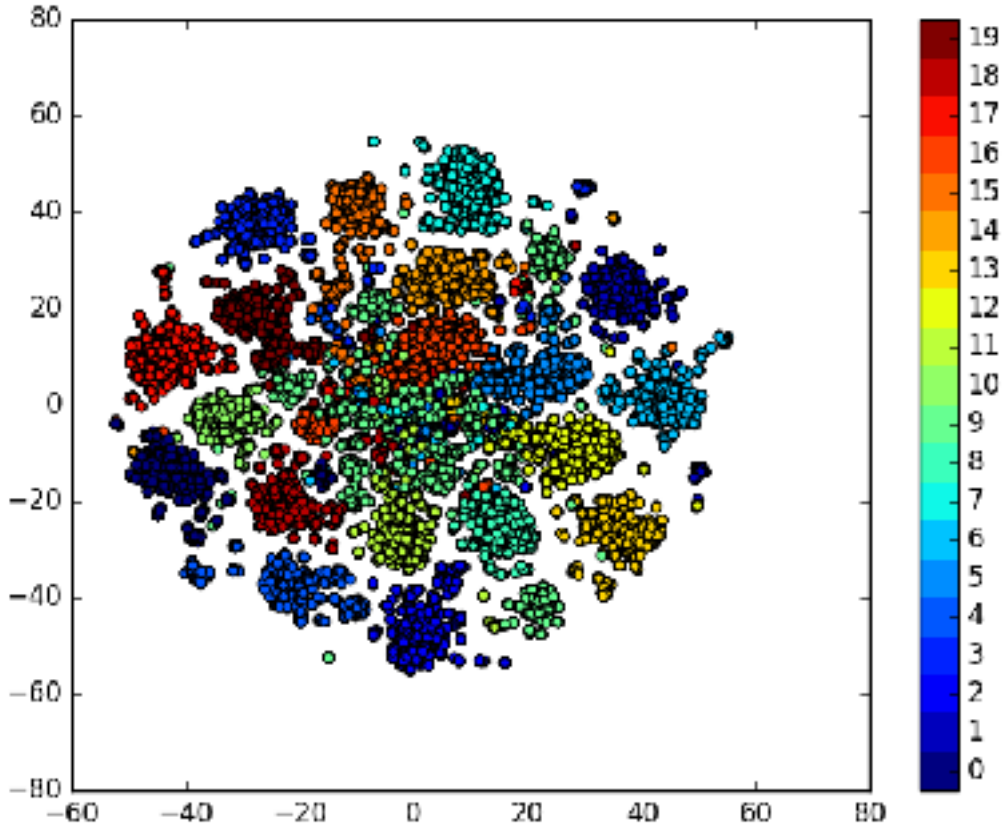
ML HW4 b02902096 王浩恩

1.Analyze the most common words in the clusters：

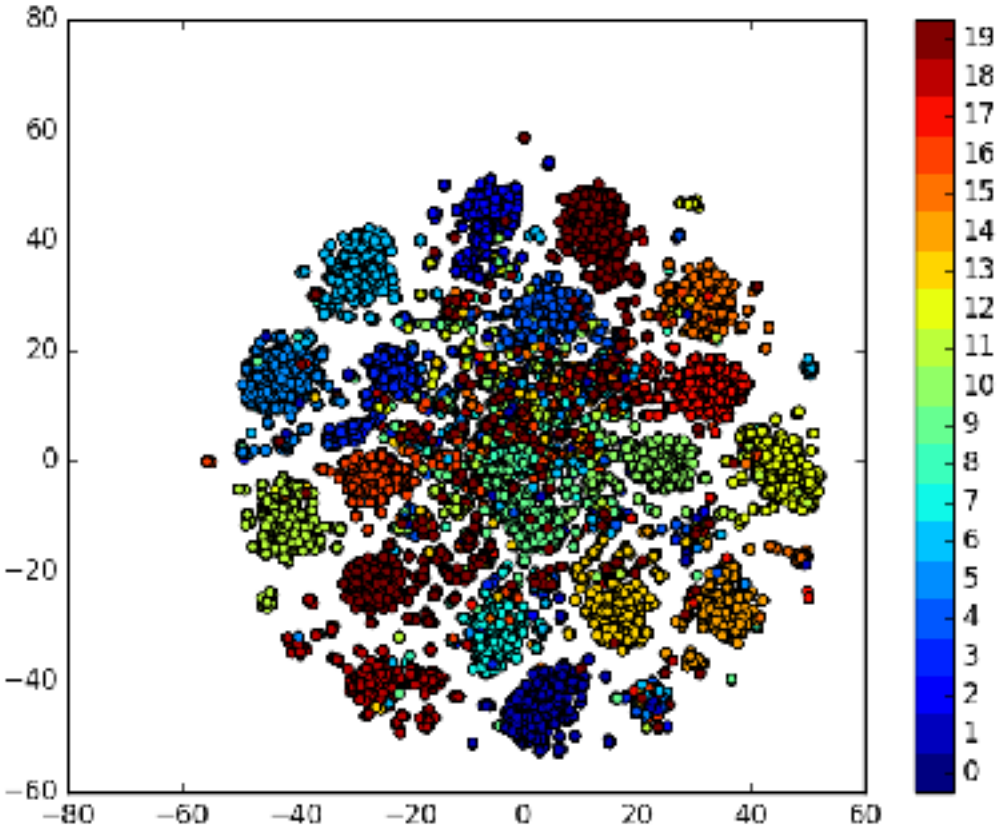
{bash, using, wordpress, haskell, excel, linq, drupalqt, visual, spring, sharepoint, apache, oracle, scala, ajax, matlab, mac os, magento, hibernate.}

these are the top-1 common words in 20 clusters.

2.Visualization:



True Label



Comments: The distributions of my cluster and true label are similar, only with some bias which leads to the shift of whole picture. Also the color of same clusters are different, but that doesn't matter for clustering.

3.Compare different feature extraction methods:

TF-IDF + LSA	BOW + LSA	TF-IDF + PCA	BOW + PCA
0.86085	0.84280	0.83884	0.78562

4.Compare different cluster numbers:

clusters	20	50	60	70	100
Kaggle Score	0.56502	0.77735	0.84987	0.85486	0.86085