Machine Learning

Homework 4

B02901143 楊筑雅
2016/12/09

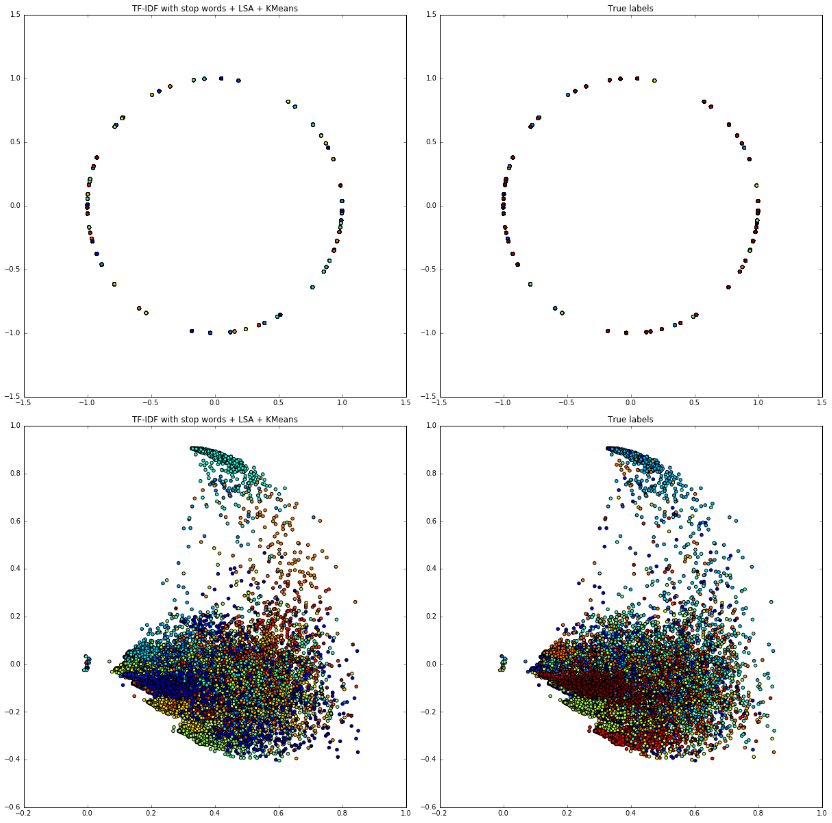
I. Analyze the most common words in the clusters

First, I use TF-IDF with stop words. Second, I use LSA to divide them into 20 groups. Last, I use KMeans to divde them into 70 clusters. Here are the most common words in those 70 clusters and their number of times are in ( ) :
zip(914), zero(148), zone(463), zero(554), zip(593), zoom(1133), working(351), xquery(274), zygo(1255), zoom(421), zone(1374), zippers(1210), xml(254), yield(165), zip(1094), zoom(1125), zip(350), xul(627), zombie(136), zero(284), wss(343), zip(909), xp(217), yesod(1091), xslt(398), youtube(263), xp(1211), zipcode(897), xml(137), xml(575), zend(465), zeros(288), zero(442), xmltype(600), xs(340), zsh(1174), zeroc(1302), zoom(463), zeros(1002), zeros(234), zendframeworks(484), zones(416), xl(342), zones(1068), yesno(438), zooming(1244), zk(1081), xdg_session_cookie(420), yes(189), zygo(1259), zoom(348), zen(516), yhc(509), xcode(377), zones(1177), zip(430), xp(421), xml(275), yaml(1179), xss(563), xml(242), zsh(910), xls(121), xml(111), zikula(1118), xstream(188), zooming(389), xslt(252), yes(210), yields(176)

We can see that there are no irrelevant words like “the” since we set stop words. In addition, we can tell that lots of StackOverflow questions are about “zip”, “zero” or “zoom”.

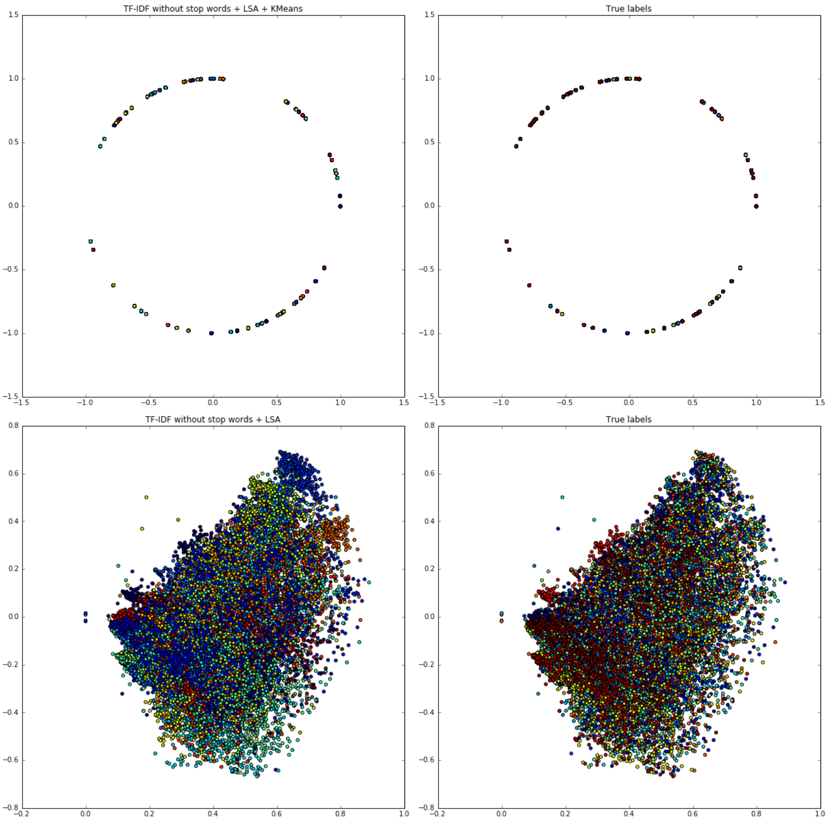
II. Visualize the data by projecting onto 2-D space

The scores are 0.84075, 0.83863 in public and private set respectively. The figures in the left side are TF-IDF with stop words + LSA (+ KMeans) , and the figures in the right side are true labels. The figures in the first row are visualized by using LSA to reduce the dimension of KMeans to 2. It is not clear, so I chose two dimensions out of 20-dimensions LSA and visualized them, which are the figures in the second row. We can see that two figures have similar distribution of dots with same color.

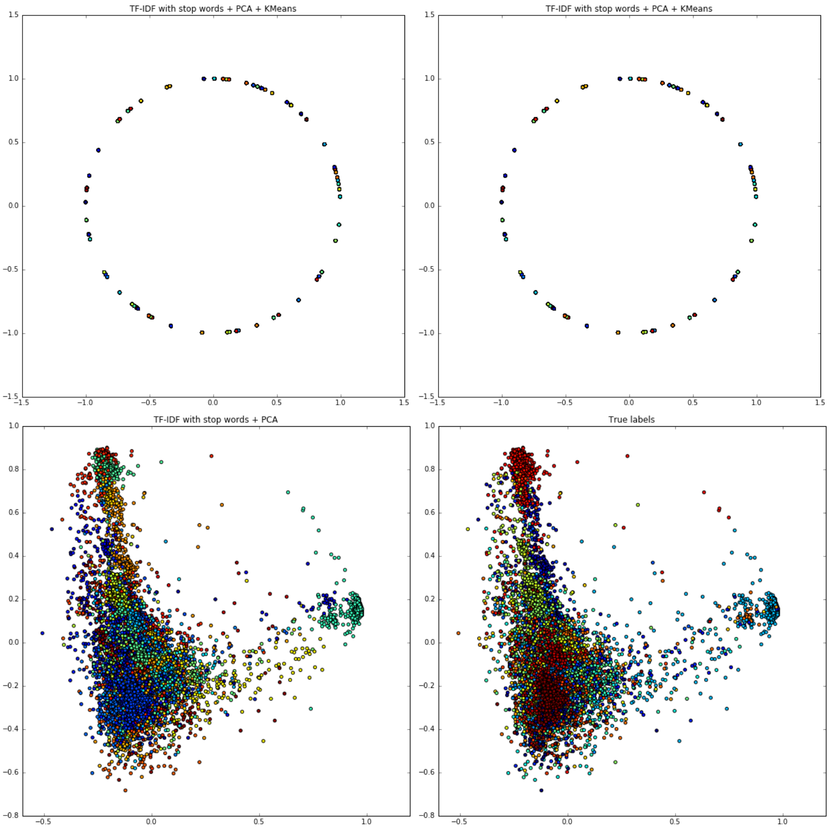


III. Compare different feature extraction methods

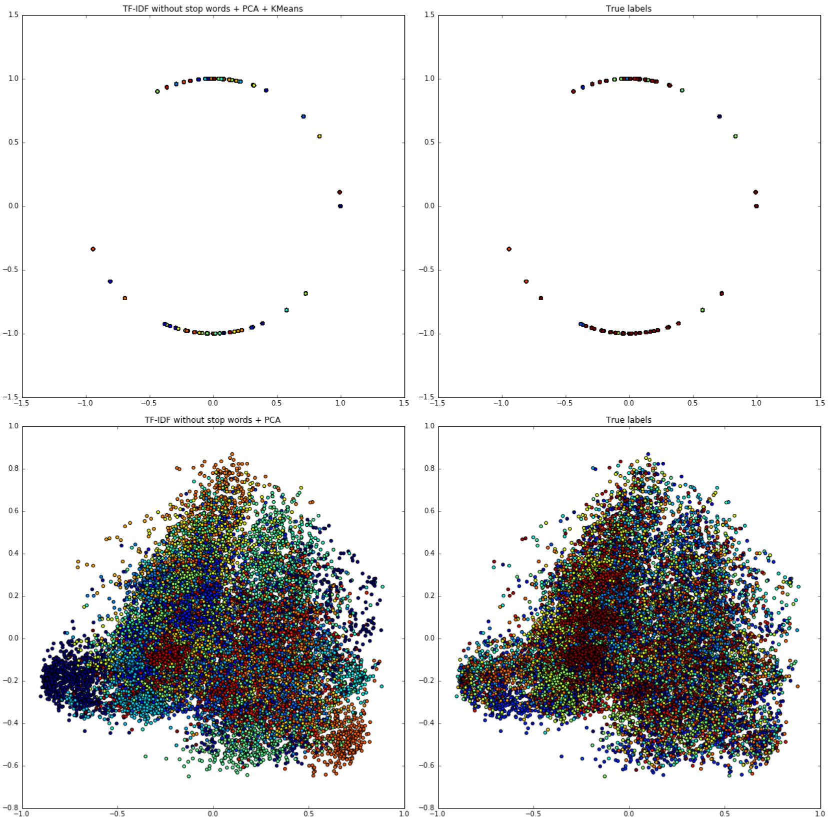
The figures below are TF-IDF without stop words + LSA (+ KMeans), and the scores are 0.50271, 0.49721 in public and private set respectively.



The figures below are TF-IDF with stop words + PCA (+ KMeans), and the scores are 0.80823, 0.80673 in public and private set respectively.



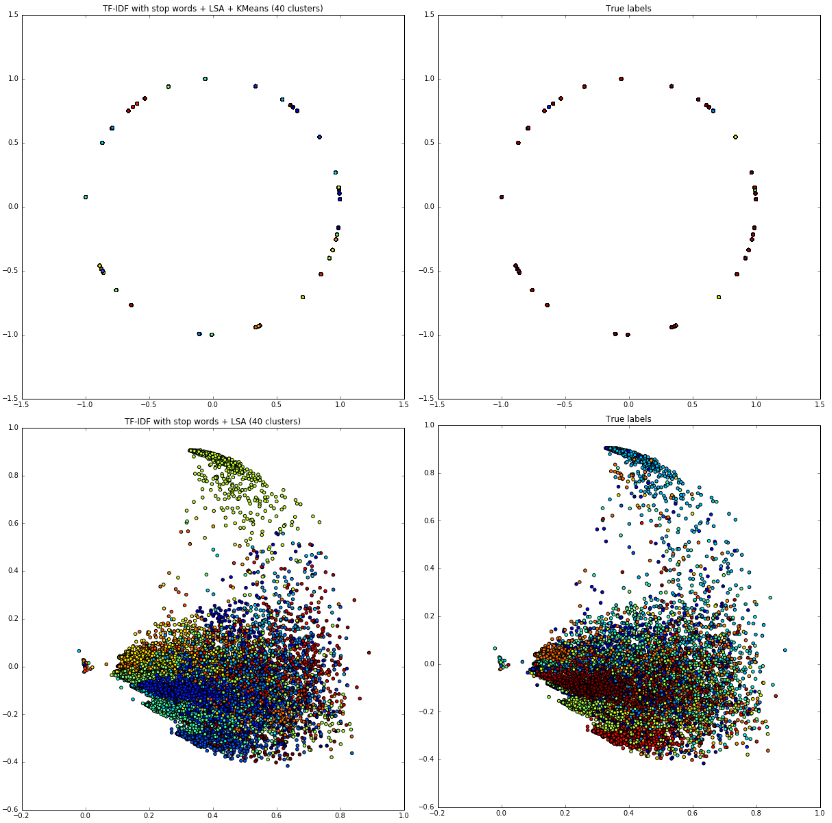
The figures below are TF-IDF without stop words + PCA (+ KMeans), and the scores are 0.43784, 0.43193 in public and private set respectively.



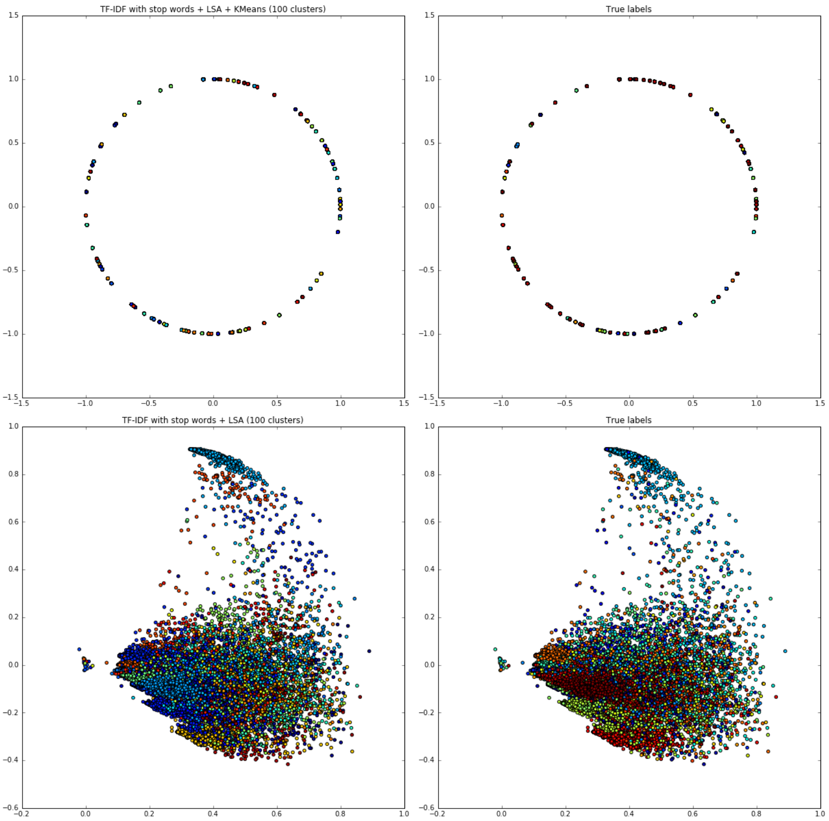
From the figures and scores, we can know that the ones without stop words has messier dots and lower performance than the ones with stop words. In addition, PCA has slightly lower performance than LSA does.

IV. Try different cluster numbers and compare them

The figures below are TF-IDF with stop words + LSA (+ KMeans) (40 clusters), and the scores are 0.81154, 0.81012 in public and private set respectively.



The figures below are TF-IDF with stop words + LSA (+ KMeans) (40 clusters), and the scores are 0.85372, 0.85294 in public and private set respectively.



From Figures, we can tell that they have similar distribution, and the performance slightly increases from 40 to 70 to 100 clusters.