Machine Learning HW4 Report

電機三 郭笛萱

B03901009

Discussed with b03901133 施順耀

1. The most common words in each cluster are shown in the following picture. In these words, some of them are labels of each cluster, but most of them are irrelevant words.

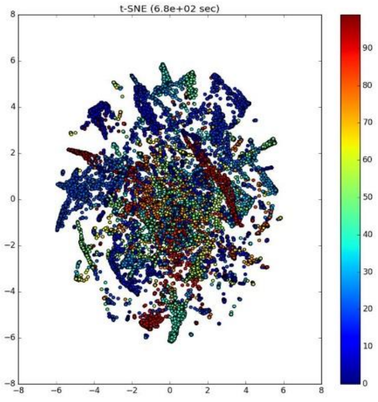
```
cluster 0 : wordpress: 872  to: 327  in: 266  a: 231  how: 204
cluster 1 : oracle: 761  to: 348  in: 284  a: 276  how: 208
cluster 2 : svn: 602  to: 362  a: 307  subversion: 283  how: 246
cluster 3 : apache: 609  to: 381  how: 170  a: 169  in: 139
cluster 4 : excel: 863  to: 379  in: 374  a: 272  how: 228
cluster 5 : matlab: 831  in: 470  a: 314  to: 313  how: 228
cluster 6 : visual: 677  studio: 649  in: 401  to: 376  a: 230
cluster 7 : a: 336  in: 312  cocoa: 310  to: 296  how: 237
cluster 8 : mac: 438  to: 357  os: 292  on: 287  x: 263
cluster 9 : bash: 665  a: 447  to: 406  in: 383  how: 263
cluster 10 : spring: 818  to: 278  in: 244  a: 206  how: 169
cluster 11 : hibernate: 859  to: 318  in: 242  a: 207  how: 165
cluster 12 : scala: 811  in: 366  to: 288  a: 281  how: 185
cluster 13 : sharepoint: 742  a: 356  in: 290  to: 284  how: 190
cluster 14 : ajax: 733  to: 258  a: 174  in: 166  how: 160
cluster 15 : qt: 627  to: 312  in: 297  a: 283  how: 216
cluster 16 : drupal: 851  to: 316  in: 294  a: 253  how: 181
cluster 17 : linq: 858  to: 468  a: 284  in: 227  how: 185
cluster 18 : haskell: 724  in: 410  a: 254  to: 214  how: 165
cluster 19 : magento: 880  in: 345  to: 286  how: 171  a: 152
```

If a word is an irrelevant word, than its Term frequency (TF) should be high, Inverse-document Frequency (IDF) should be low, and TF-IDF = TF * IDF should be low. So, I adjust the max document frequency and min document frequency of TfidfVectorizer function to get rid of those words with DF > 0.4 (proportion of the whole document) & DF < 2. Before using TF-IDF, the public score is 0.20, private score is 0.20, after removing these words using TF-IDF, the public score is 0.467, private score is 0.466, which shows a great improvement by deleting irrelevant words.

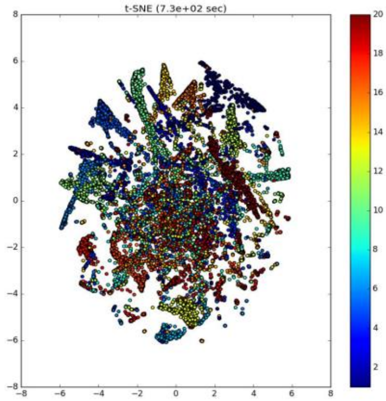
In my best model, I use *stop_words.=‘english’* in my program, which removes the words in the stop_word list provided by sklearn, and the public score is 0.885, private score is 0.883.

So, by merely removing the stop words, the result improved 0.68.

2.



My label



True label

I used TSNE to plot these two figures. The left figure is plotted by labels produced in my program, and from this figure we can see that blue and red points at the edges gathered together. The right figure is colored by true labels. We can see that more points are gathered together than the left figure. So from these two figures, I found that there exits some differences between the labels produced by the model and the true labels, and that's why the score of this model is only 0.83, not good enough.

3. The first feature extraction method I used is TF-IDF. I set max_df = 0.4, min_df = 2. With TF-IDF, the private score I got is 0.343, public score is 0.343.

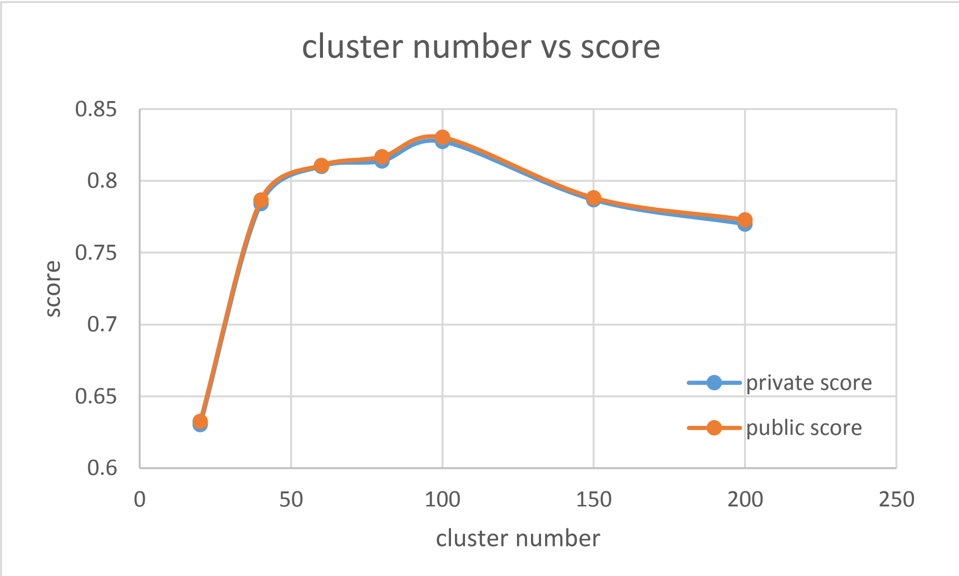
The second method I use is LSA. I have tried different n_component, and found that n_component = 20 gives the best results. With LSA (x_component = 20), the private score = 0.466, and public score 0.467.

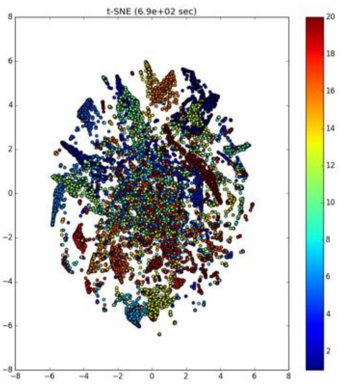
The third method I used is removing stopwords. By using the stopword list provided by sklearn, the private score I got is 0.538, the private score I got is 0.541. Which shows that removing stopwords improves the results a lot.

The forth method I use is Bag-of-word. This method count the number of different words and store these numbers as features. The private score is 0.308, and the public score is 0.310. I think the reason why Bag of word doesn't give a good result is that it doesn't get rid of those irrelevant words.

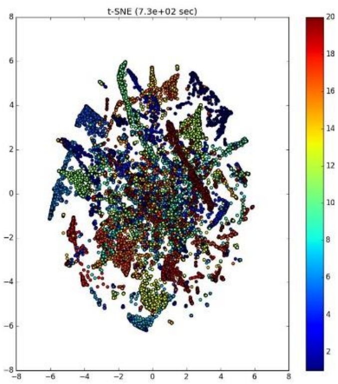
In my best model, I also use two kinds of stem, Porter stem and Lancaster stem. With these two kinds of stem and LSA + TF-IDF + removing stopwords, the result of public score becomes 0.885, and the result of private score becomes 0.883.

4. The figure below shows the public score and private score while using different cluster numbers. The results show that cluster number affects the scores a lot. From cluster number =20 to cluster number =40, the score improves almost 0.15. Cluster number = 100 gives the best public score and private score, which are 0.83 & 0.827 respectively.

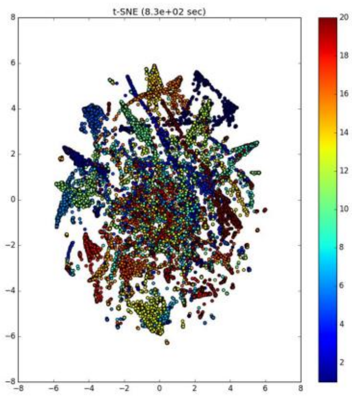




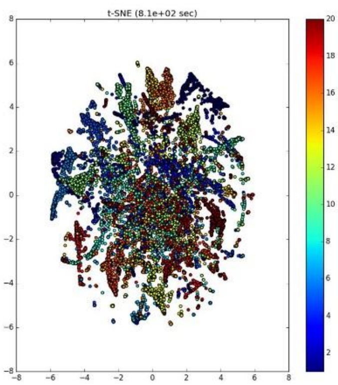
Cluster = 20



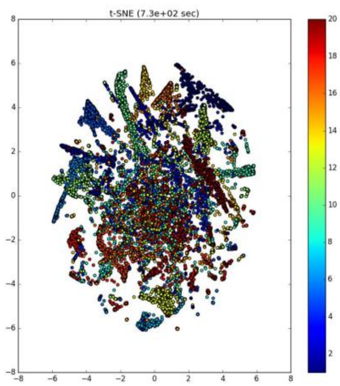
Cluster = 40



Cluster = 60



Cluster = 80



Cluster = 100