HW4

Machine Learning, 2016 Fall

R05921075, 電機一, 鄭凱文

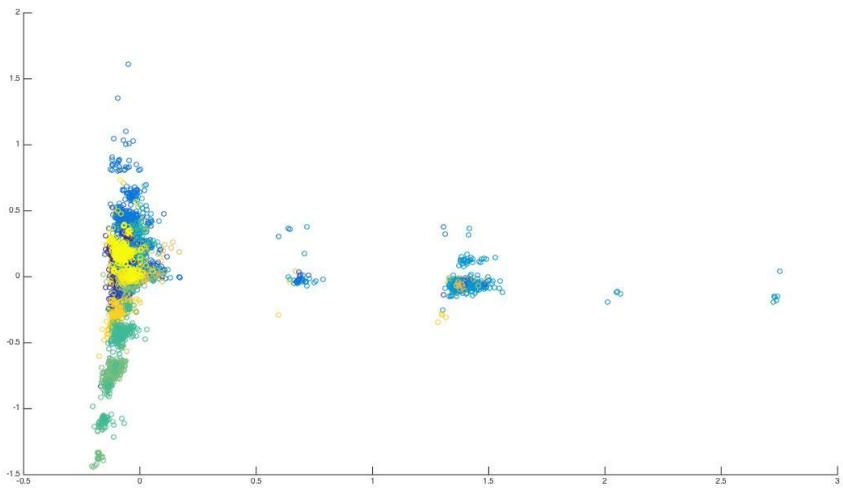
- Analyze the most common words in the clusters. Use TF-IDF to remove irrelevant words such as “the”.

I didn’t implement TF-IDF. Instead, we know there are 1000 titles in each tag, so we can naively delete the word appear more than 1000 times. It make sense that the word appear more than 1000 will confuse our classification.
Moreover, I use the toolkit **CountVectorizer** can remove english stop word.

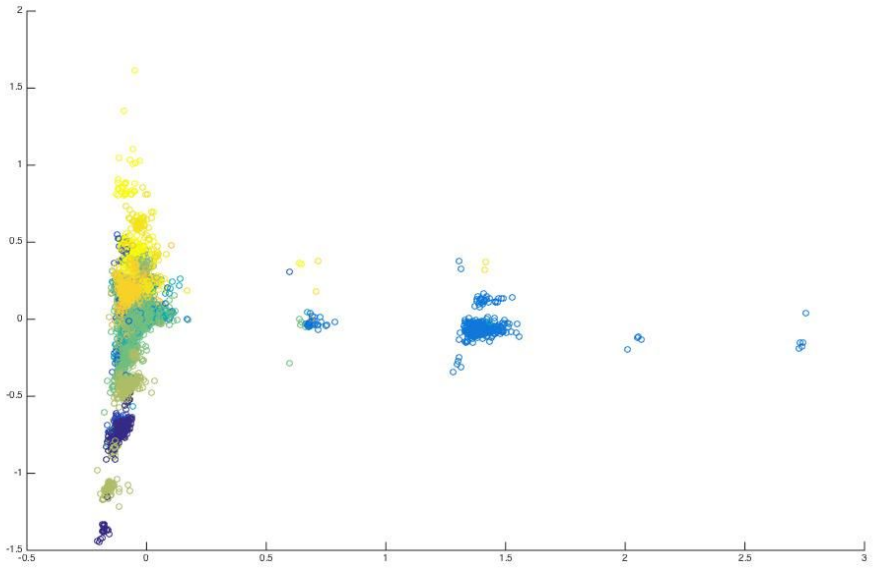
- Visualize the data by projecting onto 2-D space. Plot the results and color the data points using your cluster predictions. Comment on your plot. Now plot the results and color the data points using the true labels. Comment on this plot.

Use Matlab command `scatter(x1,x2,[],y)` to plot.

`x1`: PCA component 1, (20000,1)
`x2`: PCA component 2, (20000,1)
`[]`: default color list
`y` : label, (20000,1)



↑ True label plot



↑ My cluster label plot

- Compare different feature extraction methods.

The toolkit I used is **CountVectorizer**

```
from sklearn.feature_extraction.text import CountVectorizer
```

Method 1: Without removing stop word

```
1 trans=CountVectorizer(token_pattern=r'\b[a-z]{2,10}\b')
2 Xtmp=trans.fit_transform(txt)
3 wordnumraw=len(Xtmp)
4 for i in range(wordnumraw):
5     tmp=sum(Xtmp[i])
6     if tmp==1 or tmp>1000:
7           continue
8     else:
9           X.append(Xtmp[i])
```

Line 1 define a extractor it convert every alphabet to lowercase and extract the words length between 2 to 10. Line 2 transform the title into word matrix. Line 3 to 9 delete the words appear only once or more than 1000. Appear once cannot give any information, and appear more than 1000 will confuse the classification, because of only 1000 titles in one class.
The shape of X is (20000, 4631).
Moreover, I use a trick to make a more presice prediction. Because I find some cluster are bigger than others. As a result, it should be divided. Our task is distinguishing whether two titles are in the same class. After Kmeans determine two title in the same class, I take a inner product on two title vectors. If the value bigger than one, and I can assure they are in the same class. Other situations all see as different class.
Kaggle score: 0.75805 on public; 0.75484 on private

Method 2: Removing stop word

```
trans=CountVectorizer(token_pattern=r'\b[a-z]{2,10}\b',stop_word='english')
```

Replace Method 1 line 1 into above code.

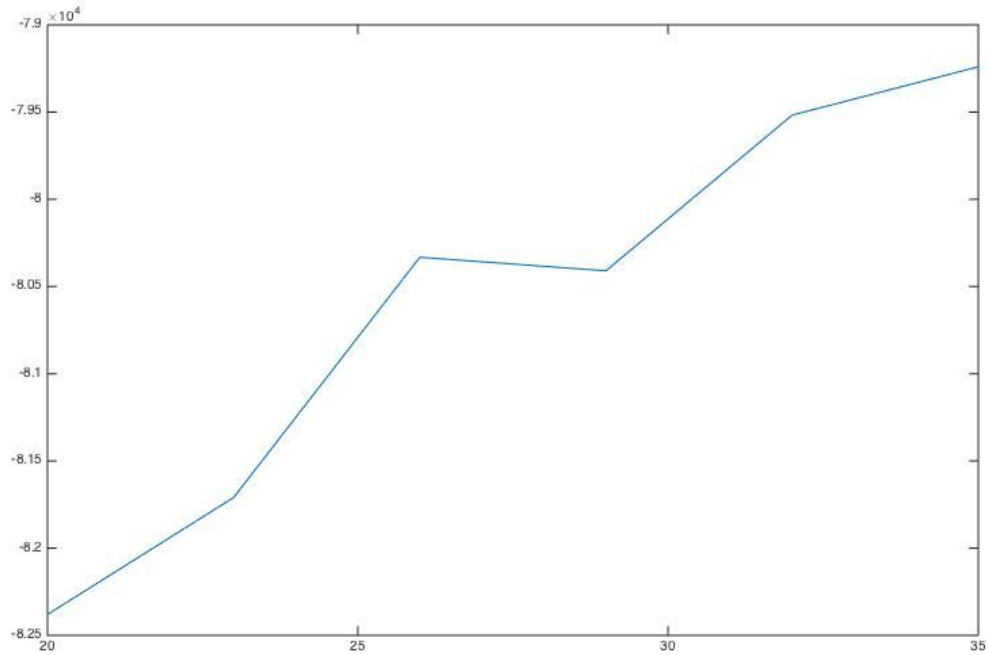
The shape of X is (20000, 4430).

Kaggle score: 0.79307 on public; 0.79033 on private

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

Use `KMeans.score()` print different cluster score.

cluster 20,-82379.7648958
cluster 23,-81709.5796475
cluster 26,-80332.3300814
cluster 29,-80409.9403661
cluster 32,-79517.3016885
cluster 35,-79240.5559886



We can find score of cluster 29 down, so divide 29 clusters may be a better choice.