ML2016 HW4 Report

R05921040 謝友恆

Problem 1: Analyze the frequent words in each cluster

Cluster number is set to 20. TF-IDF would choose the words with frequent between 2 and 80%. For readability, I didn't use stemming or lemmatization in this question. Words in cluster are ordered by frequency.

Cluster 0: magento custom product add products page get problem category order
magento is a purchase scheme, involving choosing products of different categories on different pages and with lots of custom design.

Cluster 1: haskell type function list error use problem scala string data
haskell is a language to describe function; type, list, scala are common words in it.

Cluster 2: excel using file data vba cell sheet text way macro
excel programming often uses vba to process data in file. Sheet, cell, macro are common words.

Cluster 3: use ajax qt spring oracle scala mac get excel file
this cluster have words from many tags, we may fail in this one.

Cluster 4: scala use type java using function class list way string
scala are often compared with java, since its object oriented programming, function and class would be frequent.

Cluster 5: linq sql query using multiple use get group select list
linq is a sql language, used to query database with group, select commands.

Cluster 6: wordpress page post posts plugin category custom get blog problem
wordpress is a blog scheme, user can post different category article on different pages.

Cluster 7: matlab function array plot matrix image problem file using error
all words in this category are common for matlab programming.

Cluster 8: svn file files repository subversion directory server copy working update
the words svn(subversion), repository, server, directory are close related (by google search).

Cluster 9: spring use mvc security bean file using application web framework
spring is related to javabean, mvc and security.

Cluster 10: ajax jquery problem php page get request use call using
there are many jquery problem in ajax, often wrote in php, with some request sent and function called.

Cluster 11: hibernate query mapping one problem criteria using many table object
hibernate is a object mapping solution on database(table) mapping.

Cluster 12: sharepoint list web custom site 2007 get page services part
sharepoint is a web serveice application. A famous structure in it is list. And 2007 is a popular version.

Cluster 13: apache rewrite server problem use url mod_rewrite redirect htaccess using
apache is a famous server. Rewrite function can redirect url.

Cluster 14: oracle sql use table query database way error best data
oracle is a famous database system, sql language can be used to query data in tables

Cluster 15: qt application window windows problem widget get using creator create
qt creator is a ide to develop windows applications. There are many widgets in it.

Cluster 16: bash script file command line files shell variable string get
bash(shell) script often used to invoke some commands and processing files.

Cluster 17: visual studio 2008 project 2005 files add code build use
visual studio builds codes from files in projects. 2008 and 2005 are famous versions.

Cluster 18: mac os application cocoa get way development osx best web
mac os is called osx, there are many applications.

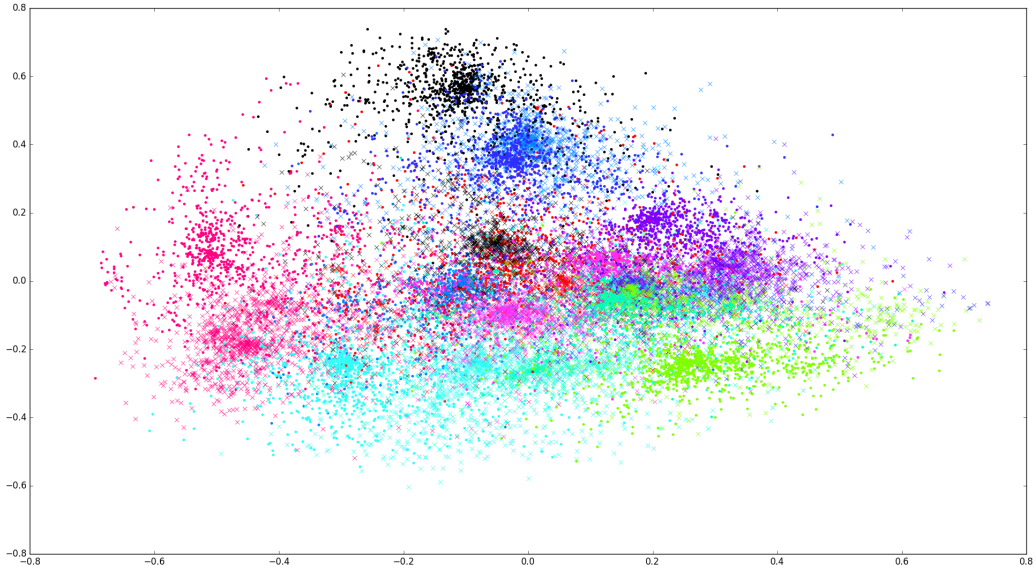
Cluster 19: drupal custom form module node content views page view add
all words in this cluster has something to do with drupal (by google search)

Conclusion: words like "use" exist in many clusters, should be added into stopwords.

Problem 2: visualize the clusters of data

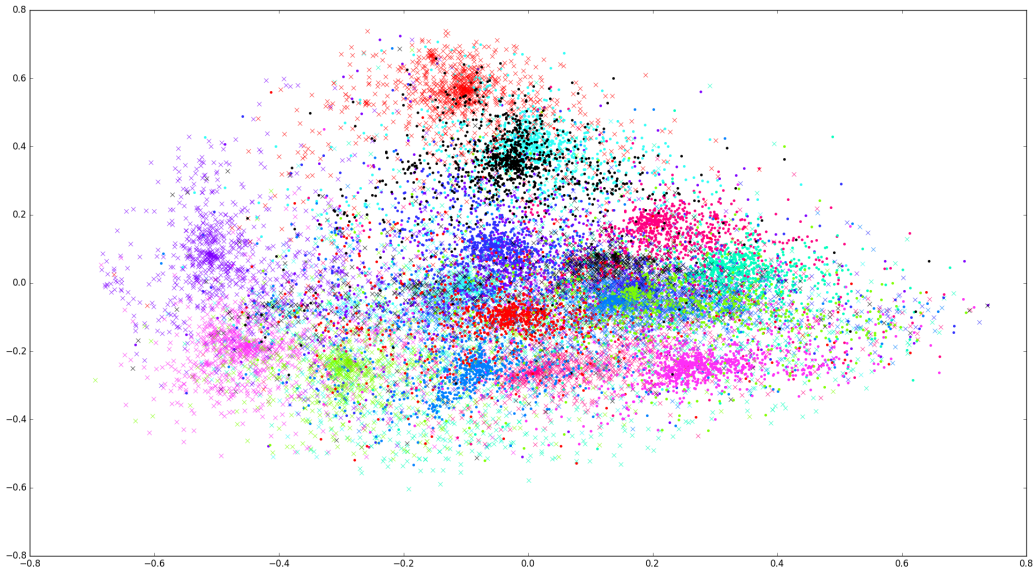
I first use lsa to reduce the feature dimension to 20, then cluster into 20 clusters using kmeans. For visualization, I further reduce the feature dimension to 2 using PCA. There're 10 kinds of colors, 2 notations ('x' and '.') for each color.

My clustering result:



The clusters in this figure overlapped with each other, this is because the data are not separated that much. In problem one, we can see some clusters shares common key words. Furthermore, since we cluster the data in higher dimensions, it is normal that the data cannot be separated on 2D plane.

The clustering result with original label:



Compare these 2 figure, we can see there's not much change to the distribution of the data. Which means our method have not so bad result. From this figure, we can verify it is not possible to separate data on 2D plain with our method.

Problem 3: Different ways of feature extraction.

1. TF-IDF: Term frequency constructs a dictionary contains the frequency of each words. This frequency is then reweighed by inverse document frequency, which will make some uncommon words more important by increasing their weight.

However, since we view each word as a dimension, the feature space become very large (about 4000). Thus harder for kmeans to cluster and increase the convergence time.

This method scores 0.17609 with 100(default) iteration number, the score increase to 0.22858 with 500 iteration number. However, this result are not very good due to the number of dimensions.

2. TF-IDF with LSA: Latent semantic analysis analyze the similarity between words and reduces the feature dimension. As mentioned earlier, TF-IDF suffers from high feature dimension, thus adding LSA is expected to have a better result. The following table shows the score under different dimensions of feature. We can see that reduce dimension to around 20 improves performance a lot.

Dim	10	15	18	19	20	25	50
Score	0.459	0.558	0.585	0.597	0.603	0.587	0.461

3. Hashing: Different from TF, we hash each word into fixed dimensional space with some collisions. We would not apply LSA on this one because it would be meaning less to fix the dimensions under LSA.

Dim	10	20	50	200	500
Score	0.06470	0.07035	0.10771	0.28111	0.19487

The result of reducing dimension by hashing is not good, however, it is still better then 1.

4. The power of stemming and stop words: we'll compare TF-IDF with and without stop words and stemming. We can see from problem 1, there's some meaningless keywords, however, if we remove stop words, those useless words become more and more.

With stop words filtering: 0.603. With out stop words filtering: 0.308.

The frequent words in each cluster are full of "is", "for", "what", "how", "of". The stop words are crucial in text classification.

Problem 4: Different cluster number

In the experiment, we found that cluster number would affect final score a lot. The following is the table of different cluster numbers and their score.

Cluster	20	21	25	30	40	50	60	70	80	90
Score	0.603	0.610	0.712	0.760	0.788	0.806	0.811	0.835	0.845	0.840

There's only 20 tag, but the result become much better as the cluster number grows. This may stems from the beta value in f measure. The beta value is the weight between recall and precision. If the beta value is too low, then the score would tend to value precision. In this homework, the more cluster, the higher the precision is. Thus we can deduce from the score that the beta value in the examination formula is relatively low.