ML HW4 Report

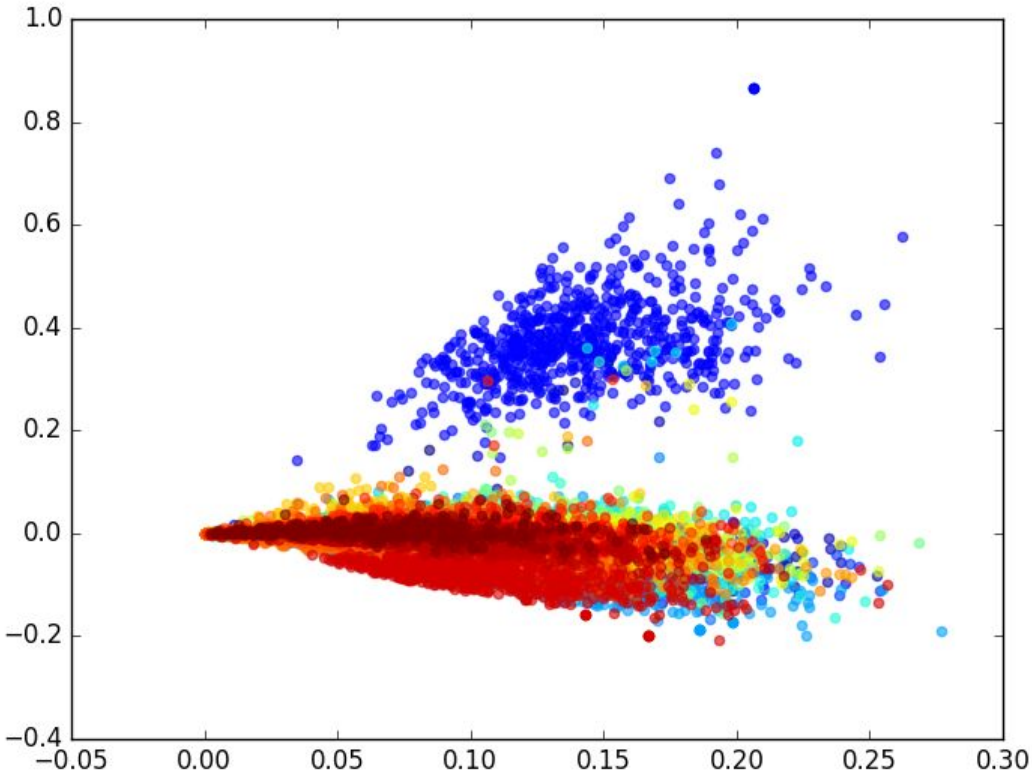
B03902082 資工三 江懿友

1. Analyze the most common words in the clusters.

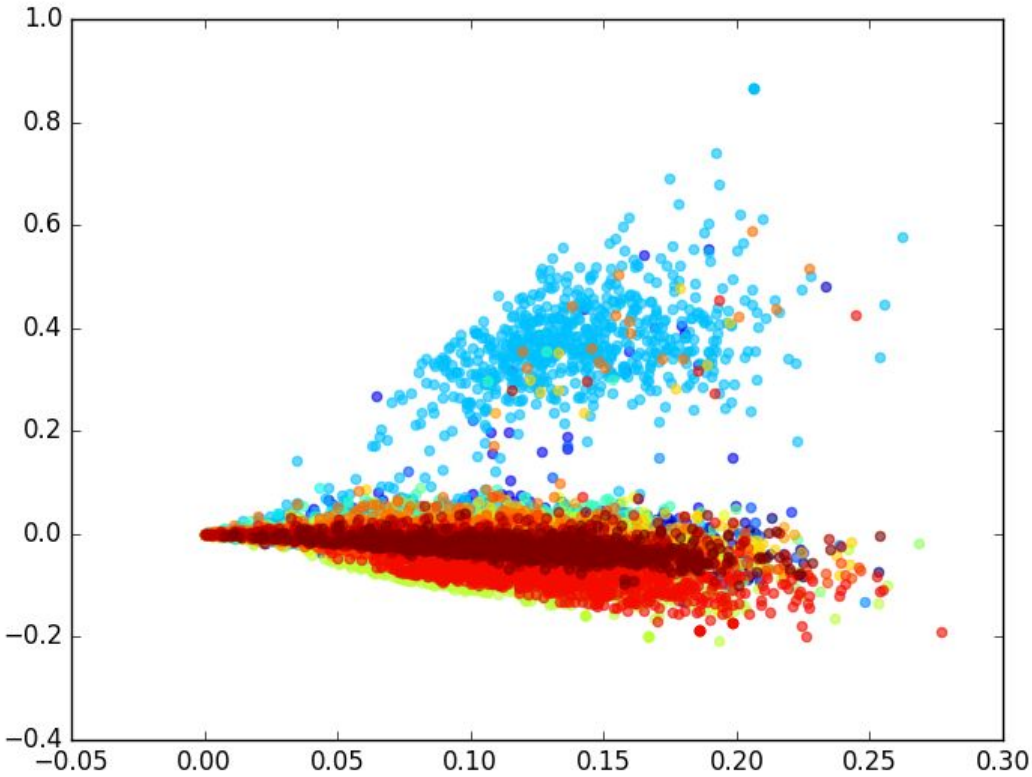
These are the words whose document-frequency is larger than 5%:
'file', 'can', 'an', 'how', 'from', 'for', 'is', 'to', 'of', 'on', 'the', 'do', 'with', 'and', 'use', 'in'

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

這是我用 mini-batch kmeans 分成 40 個團塊的結果，圖片是把 tfidf-weighted BoW vector 通過 LSA 降維到 2 維空間的結果。結果大概是分成兩個區域，上半部幾乎都是被同一個 cluster 佔據，下半部則是很多個 cluster 互相重疊，分得不是很開。



下面則是用正確的 label 把上面的圖片重新上色的結果。可以看到在正確的 label 上半部依然是被一個 cluster 佔據了，而下半部也依然是很多個 cluster 重疊在一起。



3. Compare different feature extraction methods.

Feature type	F-beta score
BoW	0.215051677629
Tfidf-weighted BoW	0.354434079027
BoW with LSA dimension reduction to dim 100	0.550722365326
Tfidf-weighted BoW with LSA dimension reduction to dim 100	0.543379023218

通過 LSA 降維前，使用 tfidf-weighted 的 BoW 效果明顯比較好；但是通過 LSA 降維後兩個的差距就變得差不多了，事實上根據 kmeans 的起始點不同有的時候是 BoW 略佔上風、有時候是 tfidf。我猜這可能是因為我們的 corpus 每個 document 都是文章標題，所以有可能每個單字的重要度都不低，所以 tfidf 提供的資訊其實不多，甚至通過 LSA 降維後 tfidf 的資訊幾乎消失了。另外我也試過用 PCA 降維，但是在這裡幾乎都是用 LSA 效果比較好。

4. Try different cluster numbers and compare them.

Number of cluster	F-beta score
20	0.310999226144
40	0.543379023218
60	0.532010811105
80	0.501961961356
100	0.465436788705