機器學習 作業四

Machine Learning HW4

R05943040 電子一 林家禾

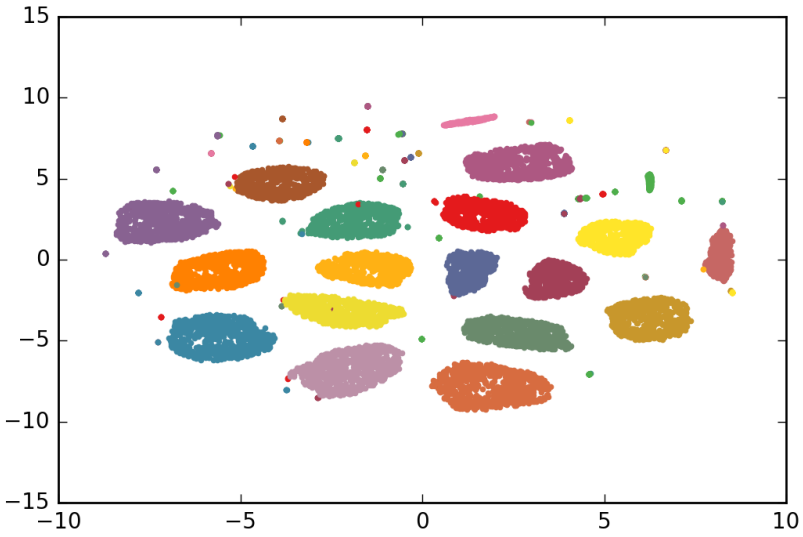
1. Analyze the most common words in the clusters. Use TF-IDF to remove irrelevant words such as “the”. (1%)

TOP50：

to,in,a,how,the,of,with,and,i,for,is,on,from,do,using,an,can,hibernate,what,excel,word press,magento,linq,drupal,spring,not,matlab,scala,file,oracle,ajax,sharepoint,visual,ha skell,s,studio,bash,apache,qt,svn,when,get,use,or,way,mac,it,does,data,list

2. 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. (1%)

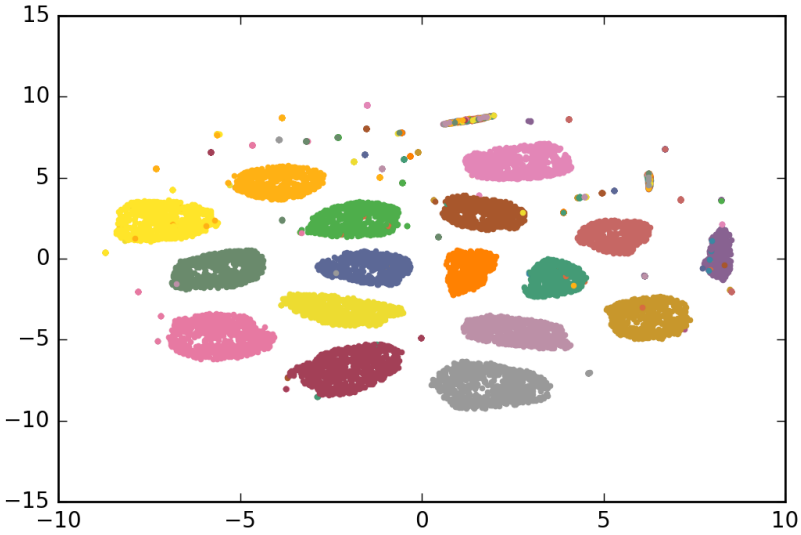
(1) my cluster predictions



成功分出 17 群，有 2 群分布僅為一直線，剩下 1 群四散在各處

四散的雜訊點推測是因為 title 中並無直接包含所屬 tag，沒有從各 title 中找出真正語意與真正的 tag 相近的詞導致分錯，也有可能是因為這些 title 同時含有多個 tag 導致誤判

(2) true labels

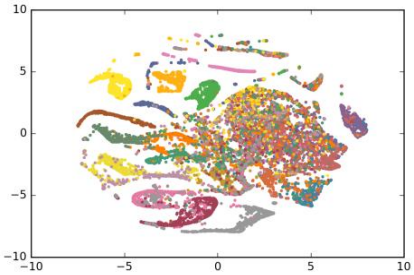


其中有兩類(圖中的直線 cluster)，並沒有成功預測真實 label，實際看這兩 cluster 自己所預測的 label 是取到了與 tag 不相關的詞句(stopwords)

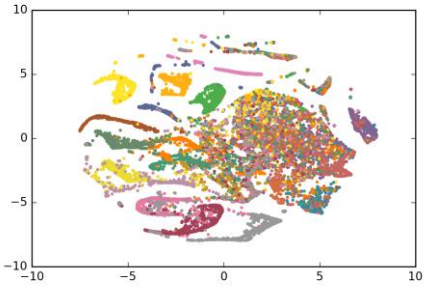
3. Compare different feature extraction methods. (2%)

自己的做法是先用 nltk 濾除 stopword，才交給 BoW、TF-IDF 處理，觀察濾除後的結果可以發現 nltk 的效果極佳，不再需要 TF-IDF 再進一步用來降低 stopword 的權重來濾除。效能變低可能是因為對 nltk 已經濾的很乾淨對 non-stopword 使用 TF-IDF 會造成反效果。導致 BoW 的效能反而還比較高

(1) BoW+PCA
Accuracy=0.50885879813893309

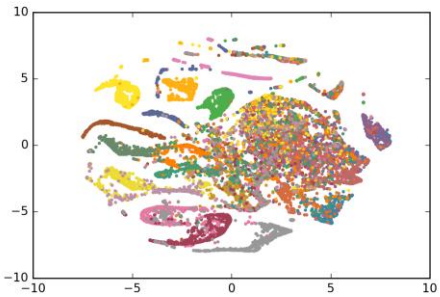
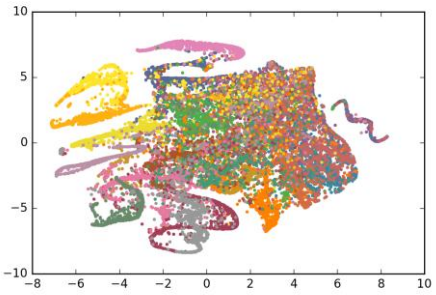


(2) BoW+LSA
Accuracy=0.52498008889815473



(3) TF-IDF+PCA
Accuracy=0.33340113250902148

(4) TF-IDF+LSA
Accuracy=0.33129134199089799

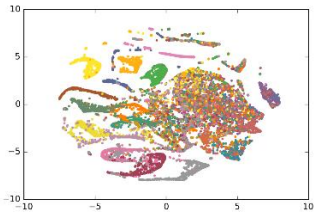
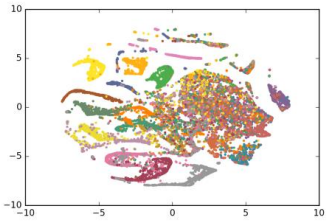
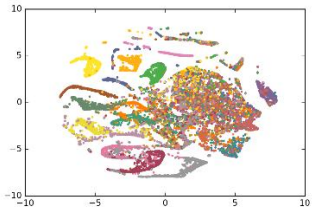
由上述的圖觀察可發現 TF-IDF apply TSNE 後的結果分布比較廣 overlap 也比較多，可能是對 non-stopword 做 weighting 反而模糊了原本字詞之間的關係

4. Try different cluster numbers and compare them. You can compare the scores and also visualize the data. (1%)

Cluster number 對 TSNE visulaize 的結果影響不大，但對 accuracy 有影響。BoW 在 100 左右效果最好，應該是在無法得知正確真實 tag 的情況下，cluster number 大於 20 才有辦法涵蓋到所有與 tag 相近的字詞。TF-IDF<50 時效果最好。可見 TF-IDF 對於濾除 stopword 之後的結果並沒有改進，所以 cluster number 越小越好

BoW

cluster numbers=50 0.405 cluster numbers=100 0.51 cluster numbers=200 0.505

IDF

cluster numbers=50 0.39 cluster numbers=100 0.33 cluster numbers=200 0.31