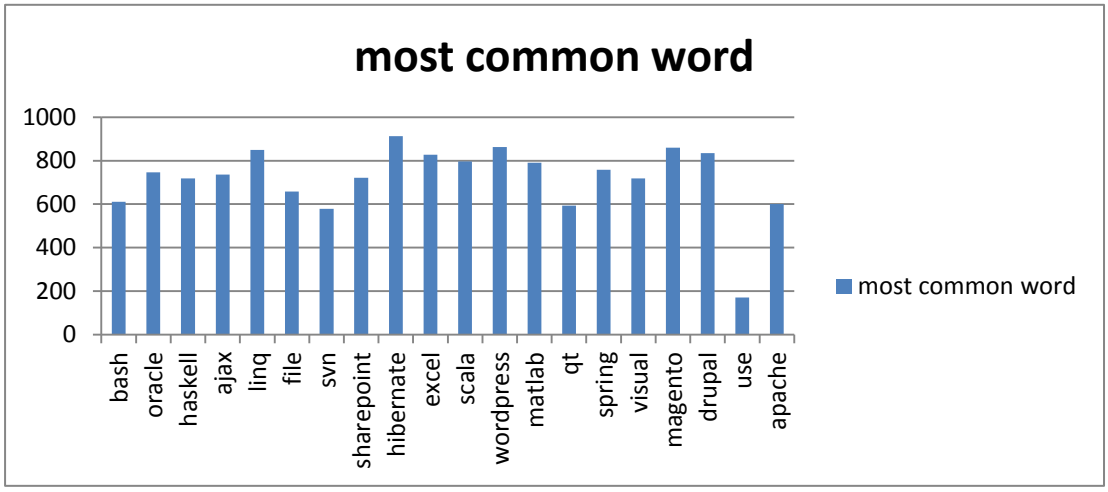
Report for machine learning assignment 4

學號: r05921012

姓名: 吳宗澤

1.

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

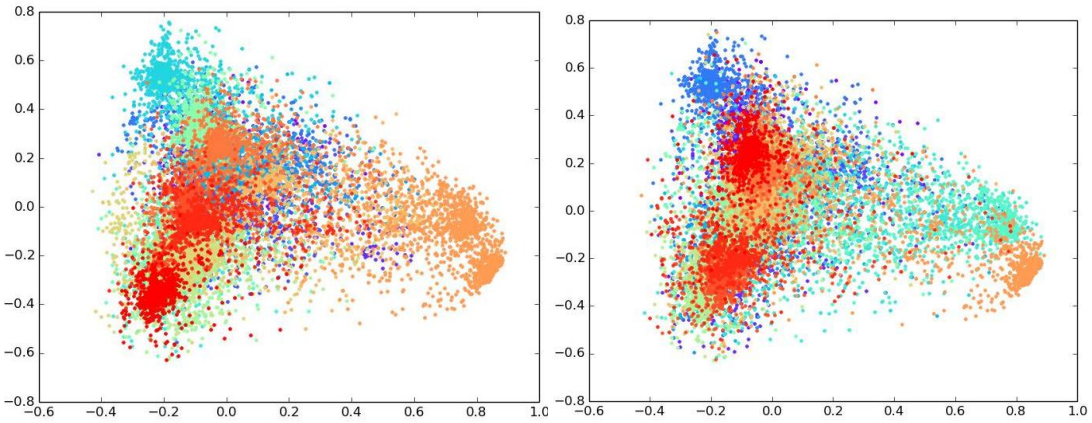


我用了網路上的 stop words list(reference:http://xpo6.com/list-of-english-stop-words/),我的架構大概是把 input data 讀進來以後作 sklearn 的 vectorizer 的動作並在同時把 stop words 給去掉，在作一個 tfidf 後丟入 truncated SVD (n_components = 20)，並利用 kmeans 去作 clustering。之後在把同一個 label 底下的在作一次 vectorizer 找出最高頻出現的詞彙。

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.

這兩張是我設 cluster = 20 的時候的圖，左邊是我自己的 prediction 右邊則是 true label 的圖，首先，因為不知道自己的分類順序是怎麼樣的，所以無法上跟 true label 裡面的一樣的代表色，但可以很明顯看出來兩者在二維投影上，都無法明顯區分出每群之間實際的分佈，可以說效果很差，但我的 prediction 在 kaggle 上可以有 85%的準確率，應該是因為 F-score 和這次評分的機制才能這麼高或是在低維度上看不出來分群的效果。



3.

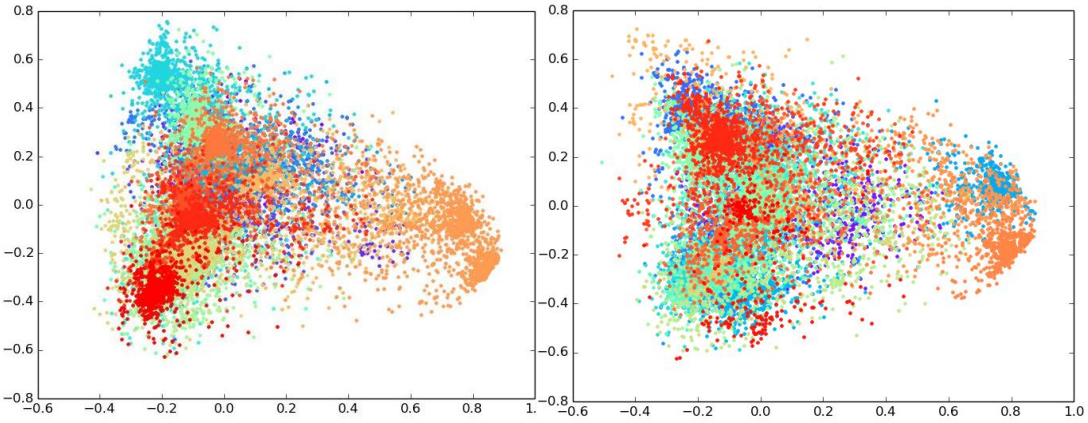
Compare different feature extraction methods.

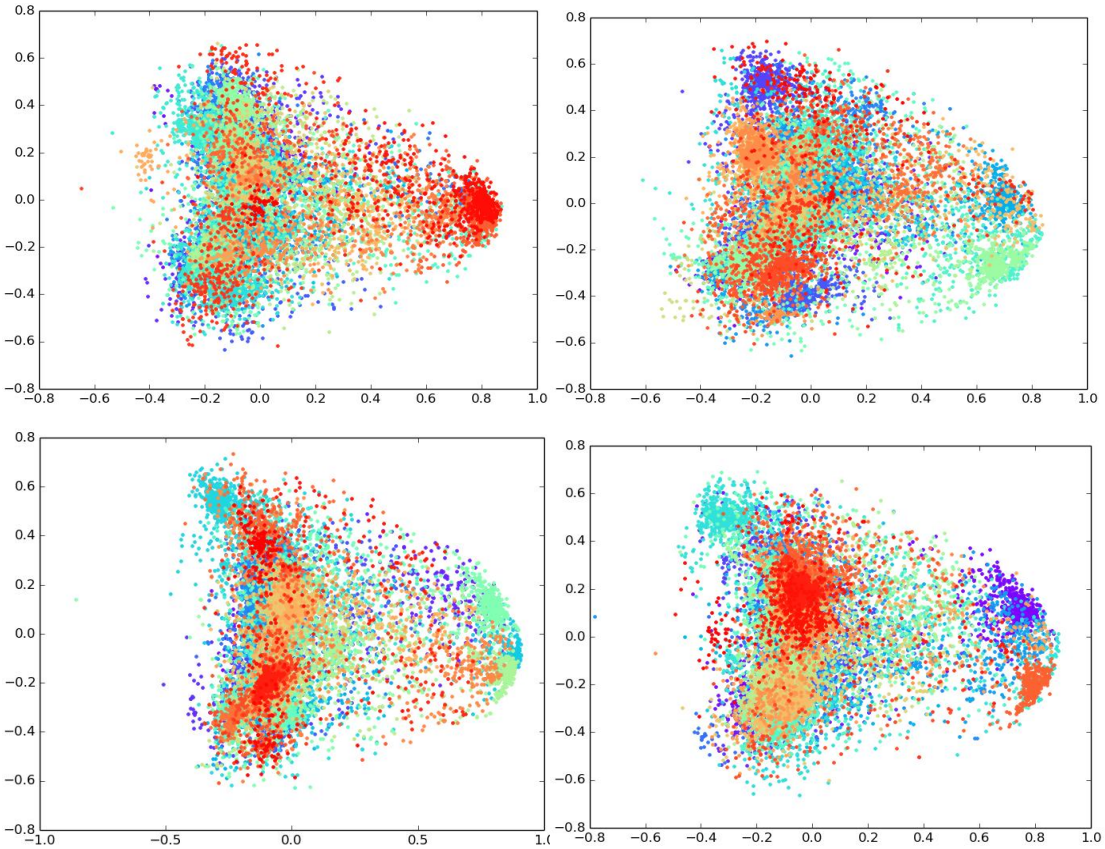
- Vectorizer+tfidf+truncatedSVD(components = 20)+kmeans(without stopwords) : 0.86588
- Vectorizer+tfidf+truncatedSVD(components = 20)+kmeans(with stopwords) : 0.56407
- Vectorizer+tfidf+truncatedSVD(components = 20)+PCA(components = 2)+kmeans(without stopwords) : 0.56407
- Vectorizer+tfidf+truncatedSVD(components = 20)+PCA(components = 2)+kmeans(with stopwords) : 0.09909

很明顯的可以發現最後的維度不能到這麼低去做區分，而且有沒有去掉 stopwords 會影響分數很大，因此還是要把它給刪除。

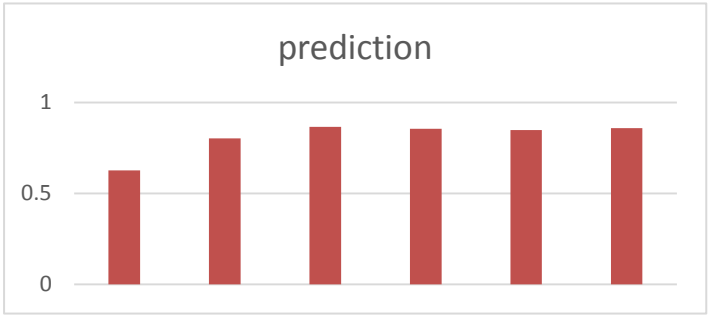
4.

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





上圖由左到右並，從上到下為 cluster 是 20,40,60,80,100,120 個
下圖由左到右分別是 cluster20,40,60,80,100,120 在 private 上的成績



可以發現在 cluster 數量為 60 的時候，有最好的表現，而在圖上還是很難分析，因為 kmeans 分類的方式好像有點隨機性，無法準確的在每一張圖去作比較