HW#4 Report

B03901027

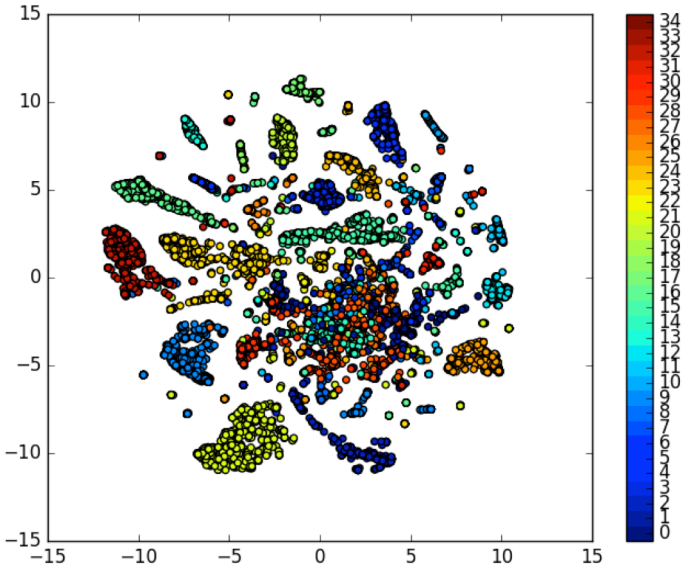
徐彥旻

ML2016, Fall

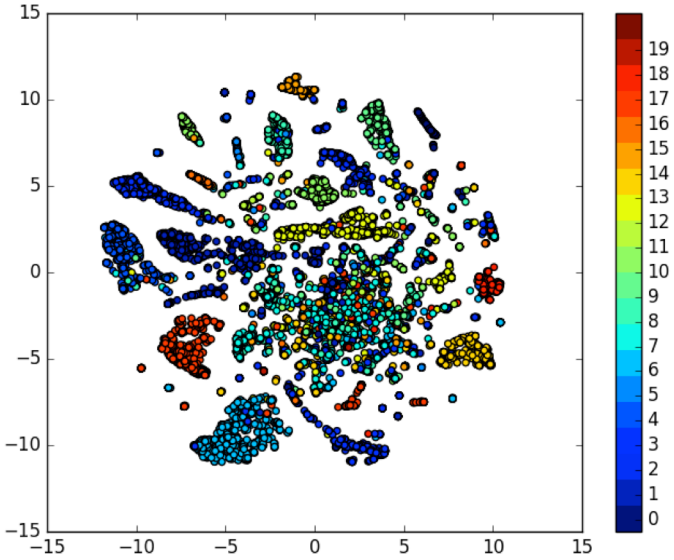
1. Analyze the most common words in the clusters. Use TF-IDF to remove irrelevant words such as “the”. (1%)

各組的字在經過 TF-IDF 處理之後，取各組前十個特徵字來分析，發現各組特徵字有所差異，但是也互有重複，像是 'using' 就出現在幾乎所有的 cluster 當中，'file'，'data' 也出現在將近一半的 cluster 中，可以考慮將這些共通的特徵字加入 stop words 當中。

2. Visualize the data by projecting onto 2-D space. (1%)



圖一 clustering result



圖二 true labels

左圖是由 TF-IDF 以及 lsa 處理，以 kmeans 分類後所得之結果；右圖是左圖以 true labels 重新著色之結果。以左圖左下為例，資料點都被 kmeans 判斷為相同的類別，對應到右圖的左下，在真實的標籤上也是同一類別，代表分類成功。整體來說，有把大部分的資料分開來，但仍有些資料是分不開的。（注意：左圖的群數設置的比真實的多）

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

1. TF-IDF + lsa + Kmeans

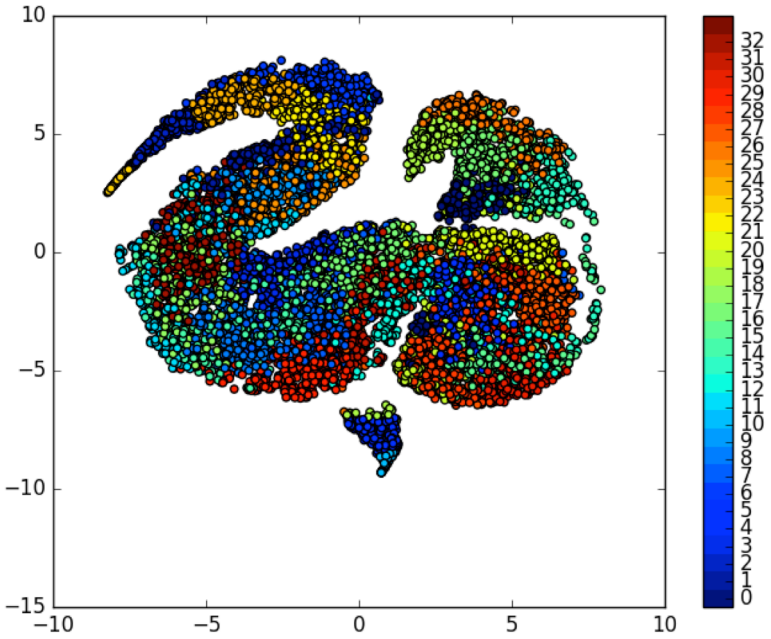
表現最好的方法，kaggle 的表現約為 0.78

2. TF-IDF + lsa + autoencoder + Kmeans

原本預期會比第一個方法表現更好，但在 kaggle 上的表現為 0.76 ，推測是 autoencoder 沒有訓練起來，可能是架構沒有調好的關係。

3. TF-IDF + autoencoder + Kmeans

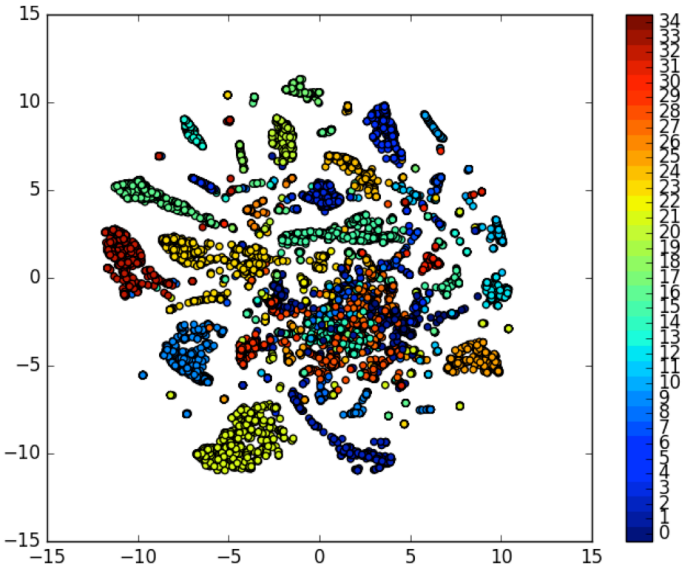
在 kaggle 上的表現不佳，幾乎無法正確預測(分數為 0.05)。



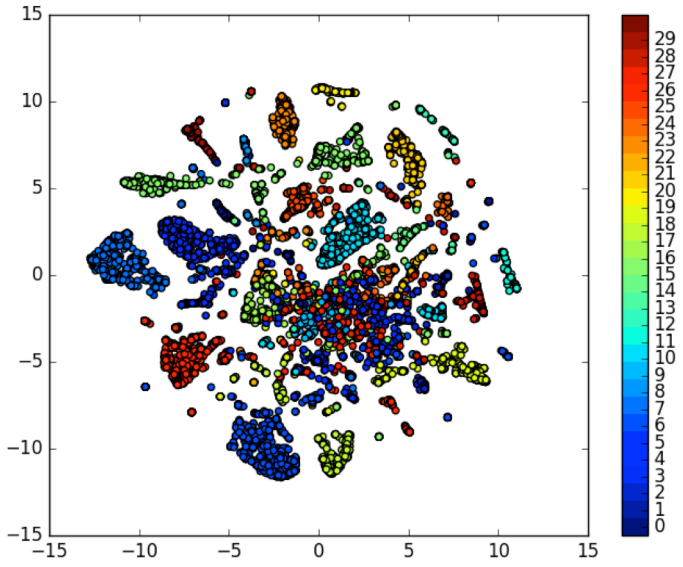
圖三 autoencoder

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

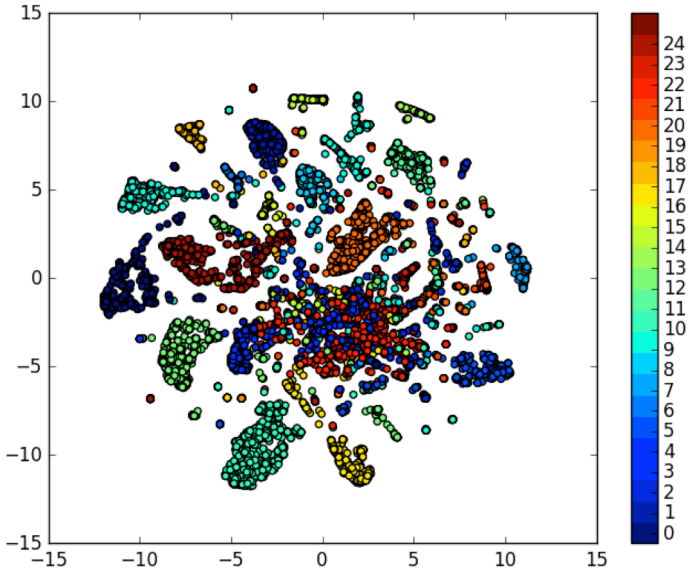
Number of clusters	performance
35	0.78
30	0.76
25	0.72



圖四 35 clusters



圖五 30 clusters



圖六 25 clusters