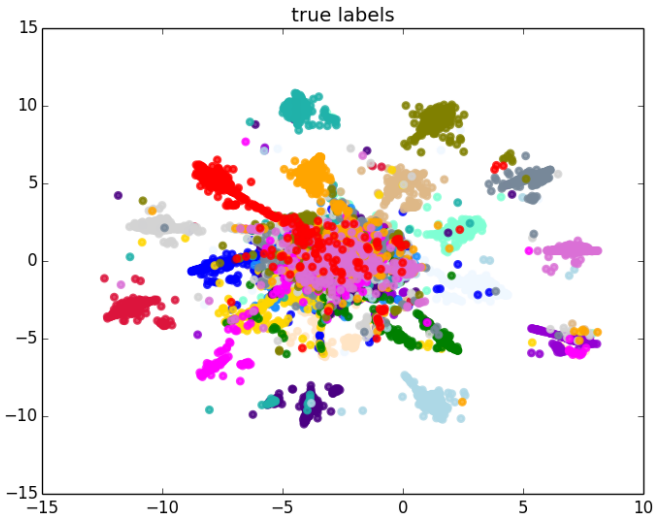
ML hw4 Report

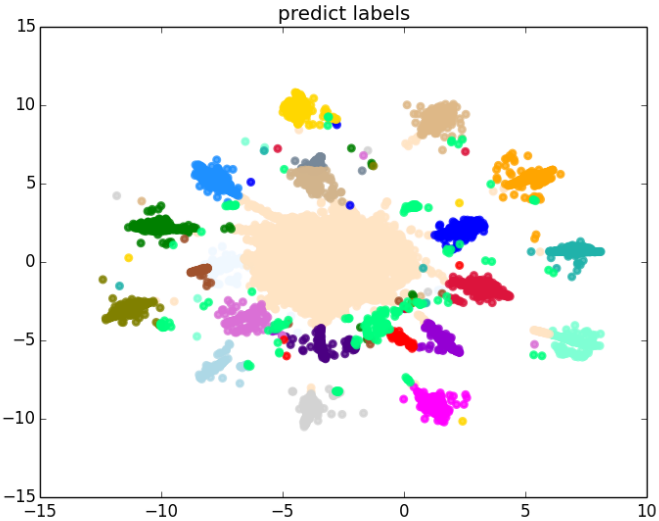
1. Analyze the most common words in the clusters. Use TF-IDF to remove irrelevant words such as “the”.

   我在網路上收尋英文字裡的 stopword，並且節錄了大約 300 個常出現的單字，當作我的 stopword.例如: ‘a’, ‘able’, ‘about’, ‘after’, ‘again’, ‘against’, ‘all’, ‘allow’, ‘also’, ‘am’, ‘an’, ‘and’, ‘any’…等等

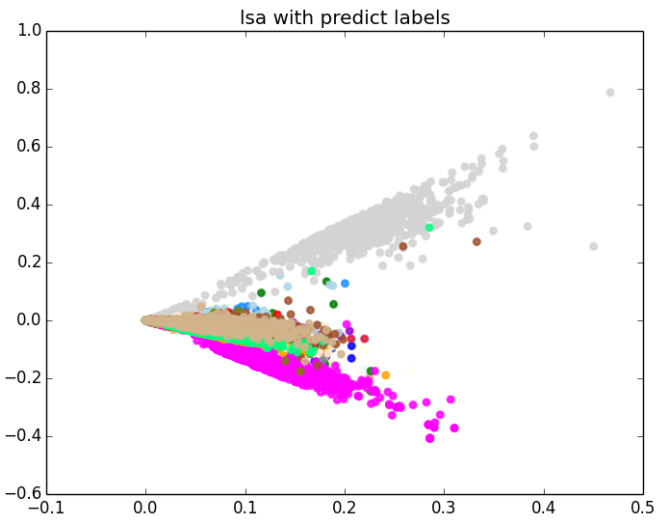
2. Visualize the data by projecting onto 2-D space. Plot the results and color the data points using your cluster predictions. Comment on your plot. Now plot the results and color the data points using the true labels. Comment on this plot.

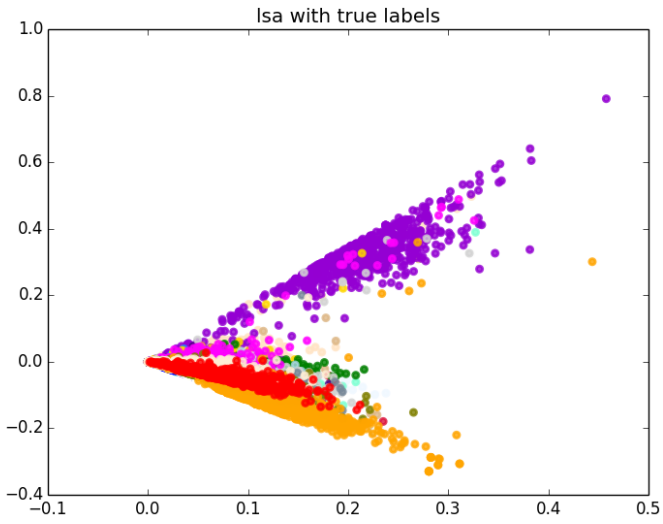
   在這裡，我採用 2 種 visualize 的方式，第一種為 t-SNE，第一張是使用真實 label 的圖，可以看到除了中間那部分並沒有把不同 label 的點分開以外，四周是可以看到明顯的分群效果的。第二張是使用經過 kmeans 後的 label，kmeans 把在真實 label 分群不明顯的 data 看成新的一大群，在圖中為米白色。比對真實 label 畫來的圖，在預測 label 的圖的四周，每一群都可以對應到真實 label 的一群，證明 data 在經過 feature extraction 後，還是保留每個 label 應有的特徵，並且在高為空間中不同 label 是分開的。





第二種方法，我使用 LSA 直接降維至 2 維，雖然也沒有產生完美的分群效果，但是可以看的出來他們在低維空間是有群聚的效應，可以猜想他們在高維空間可能會有分群的效果。





3. Compare different feature extraction methods.

在我測試的時候，我固定會把 data cluster 成 25 群。
我一剛開始使用 PCA 作為降維的工具，大約降到 100 維，但是分數不是很好，只有 0.29 剛好超過 baseline。之後把維度降到依舊沒有增加 cluster 的準確率，分數依然還是 0.29。最後我使用 lsa 降維，在經過反覆的測試，我採用 lsa 降維到 18 維會得到最好的分數大約為 0.722。

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

這裡我統一都使用 lsa 降維至 18 維，並比較不同群數之間的表現
分成 300 群 score:0.829
分成 200 群 score:0.829
分成 150 群 score:0.836
分成 125 群 score:0.838
分成 110 群 score:0.845
分成 100 群 score:0.843
分成 50 群 score:0.818
分成 20 群 score:0.709
分成 10 群 score:0.639