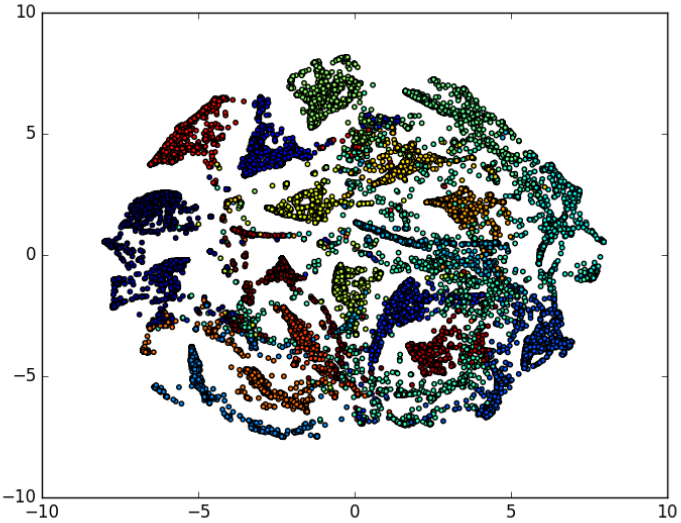
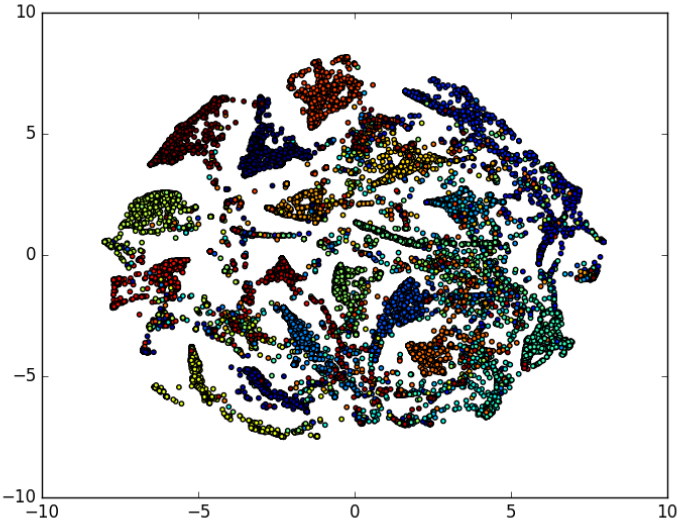
R05921063 陳定楷 HW4 討論對象：F04942066劉楚彤、R04942143吳兆倫

1. **Analyze the most common words in the cluster:**
   使用TF-IDF並刪掉stop words (sklearn預設）後，用LSA壓縮成20維，最後使用K-means分成20個 clusters，各cluster中最常出現的字依序如下：
   hibernate, linq, wordpress, excel, mac, scala, bash, svn, use, apache, drupal, spring, ajax, sharepoint, visual, oracle, matlab, magento, qt, haskell
   其中第15個cluster的visual可能為原標籤的visual-studio，而第5個cluster的mac可能是融合了 osx和cocoa兩個標籤， 第9個cluster中出現頻率最高的前十個字分別是：
   use, does, way, error, code, best, vs, multiple, make, problem
   可知此cluster可能是較基礎的程式相關問題。

2. **Visualization:**
   My prediction:



   Ture labels:



   除了圖形右下散布較雜，基本上預估與實際label相符

3. **Compare different feature extraction methods:**

   (a) BoW+Kmeans(20 clusters): Kaggle score: 0.227

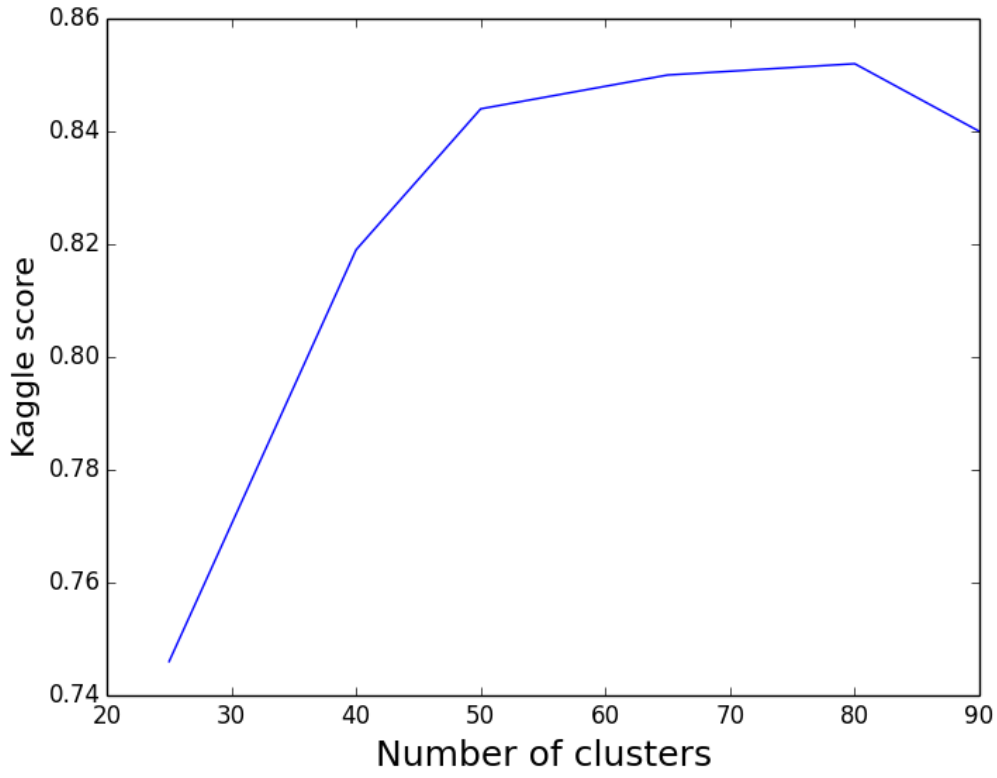
   (b) TF-IDF+Kmeans(20 clusters): Kaggle score: 0.223

   (c) BoW+LSA(dimension=20)+Kmeans(20 clusters): Kaggle score: 0.549

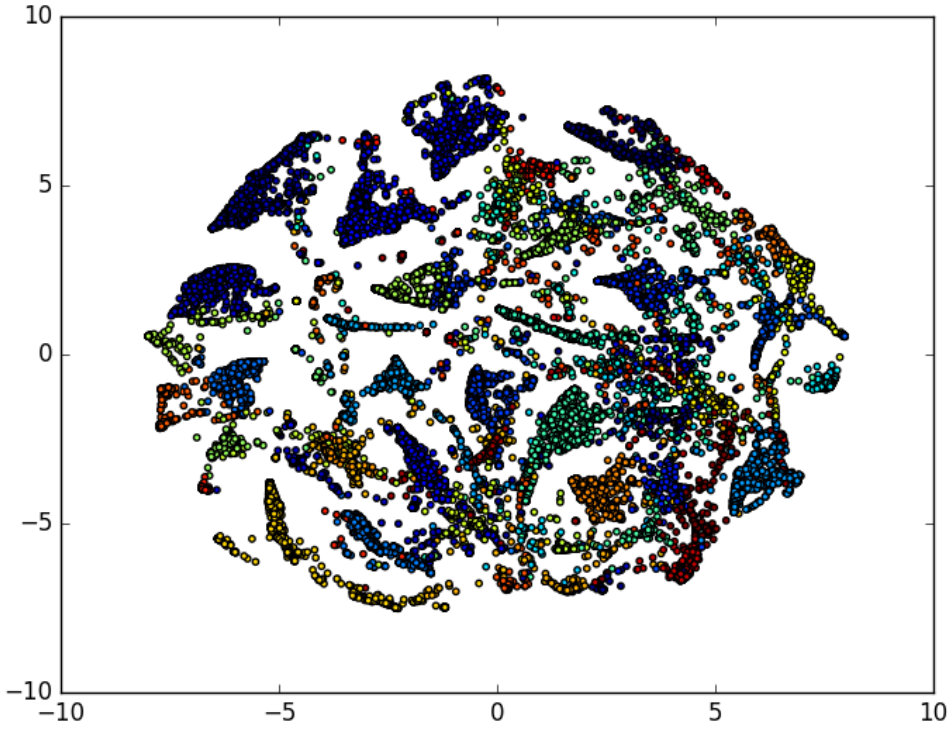
   (d) TF-IDF+LSA(dimension=20)+Kmeans(20 clusters): Kaggle score: 0.645

   正確分類效果：(d)>(c)>(a)>(b)
   可以看出當有使用LSA時，分類的效果優於沒有使用。另外TF-IDF在有使用LSA的情況下比直接使用BoW更能 顯現每個字詞的重要性。

4. **Try different cluster numbers:**



   因為$\beta = 0.25$，所以false positive對分數的傷害會比false negative來得大，因此將cluster 的個數提高可以有效降低false positive以提高分數。
   下圖為cluster = 80的預測分佈圖：



参考資料：matplotlib.org, scikit-learn.org