

Machine Learning 2016

系級：電機四

姓名：鍾勝隆

學號：B02901001

HW4 Report – Unsupervised Learning

1. Describe my method.

我使用 sklearn 中，針對文章 feature 的截取套件 TfidfVectorizer，將 Title_Stackoverflow.txt 的每個標題當作是獨立的文章，進行 vectorizing，扣掉 stop word，並使用 bigram 以增加 feature，完成 vectorization。接著用 LSA 降維，最後使用 KMeans 作分群。

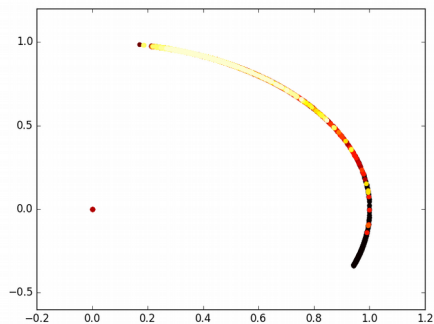
2. Analyze the most common words in the clusters.

Cluster	Word(feature)	Cluster size	Cluster	Word(feature)	Cluster size
1	excel	1053	11	sharepoint	872
2	qt	761	12	magento	944
3	svn	1144	13	visual	773
4	wordpress	958	14	apache	909
5	matlab	902	15	drupal	1002
6	use	2212	16	hibernate	931
7	scala	878	17	haskell	887
8	ajax	878	18	spring	898
9	oracle	925	19	bash	910
10	linq	939	20	mac	1224

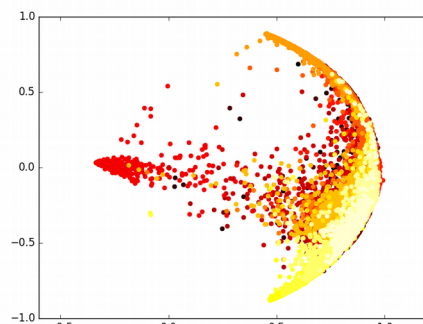
上表為使用 KMeans 分類成 20 群的結果，把這 20 群與以正確的 tag 比較，發現少了兩個 tag(osx, cocoa)，第 20 群的 mac 應該對應的是 osx，這兩個字本來就是關連性非常高的，但是 cocoa 對應的是第 6 群中最常出現的字，use，就會發現在這一群可能就有很多 Kmeans 無法正確將其進行分類的模糊資料。另外，正確的分群應是 20 群皆有 1000 筆資料，然而第六群卻是 **2.2 倍**。

3. Visualize the data by projecting onto 2-D space.

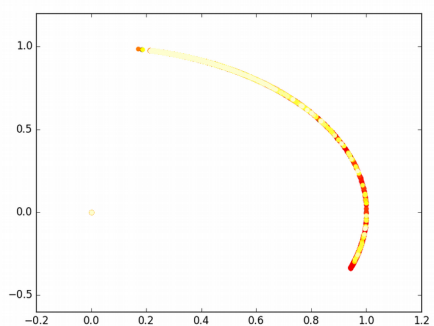
我首先使用 LSA 將維度都降到 2（圖一、圖二），但是這樣表示資料非常的不清楚。不易觀察出差異，我該改以 LSA 將維度都降到 3，在投影至 2 維的圖形（圖三、圖四）。



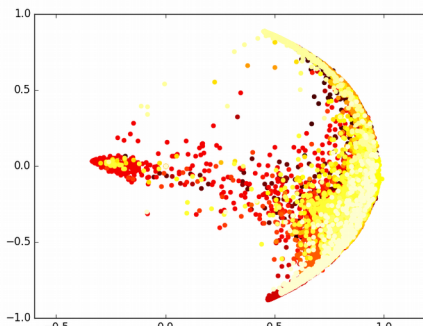
圖一



圖三



圖二



圖四

圖二和圖四皆為使用正確的 label 標出的圖，由於每次的分群不一定對應的編號相同，所以，顏色本身沒有參考性，分佈位置才需要討論。由圖三和圖四的差異可以發現，使用 KMeans 相同顏色的分佈，會比較集中，趨近圓形，但是正確的 label 分佈是較為畸形的，有很多突起。KMeans 不適合處理這種分佈，因此，造成分類的誤差。

4. Compare different feature extraction methods.

1. BoW(Bag of Words): 把所有出現的 word 都當做作 feature，維度也會非常大，容易特殊字，甚至是錯字的影響，而一般文章中，與分類確切相關的字並不一定是出現次數最多，所以單純只用 BoW 的分類的準確度一定不好。

2. Tf-idf: 將字的出現次數也考慮進來，IDF 與該字出現的文章數相關，由 sklearn 套件中設定 max_df 和 min_df，即可以將一些幾乎有出現在每一個文章中、對分類沒有幫助的字，和只出現在非常少數文章的字剔除。我的設定為超過 5 成的文章有該字，或是只有一篇文章有該字，就不考慮將其當成 feature。

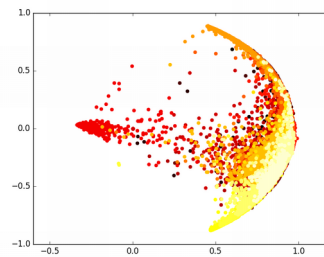
3. LSA(TruncateSVD)：降維後可以使得 KMeans 在分類上執行得更快速，減少運算時間，同時不會失去太多資料間的相對關係，但是的確資訊有減少，是增加效率的 tradeoff。

4. Exclude stop word：每個語言中都有該語言常出現的字，像是英文中的 the、and 或是 because，這些字不一定會在 Tfidf 中去除，特別用 Vectorizer 將 stop word 除去在 feature 之外。

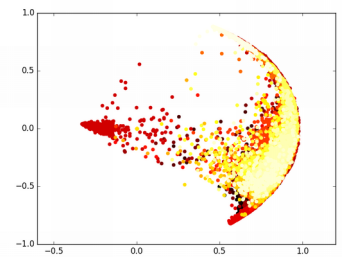
5. Try different cluster numbers.

Cluster 數量	Score on kaggle(private)
20(圖五)	0.62023
40(圖六)	0.82464
60(圖七)	0.85285
80(圖八)	0.86110
100(圖九)	0.86012
120(圖十)	0.83858

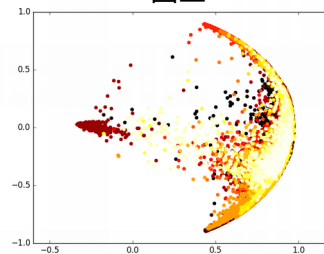
可以發現隨著 cluster 數量的增加，label 的分佈也越來越與真實的 label 分佈相近但是，由於同樣地資料會被分在不同的 cluster，所以分得過多的時候，score 反而會下降。



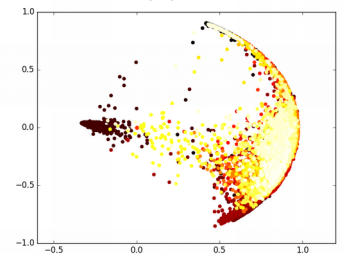
圖五



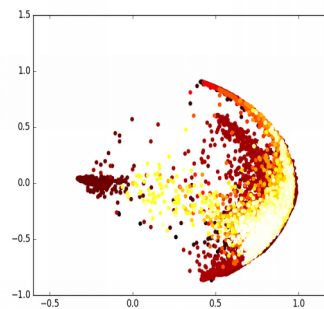
圖六



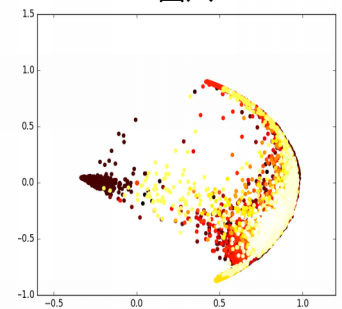
圖七



圖八



圖九



圖十