ML2016 HW4 report

姓名: 徐有慶　學號: R05922162

1. Analyze the most common words in the clusters.

Cluster	Word1	Word2	Word3	Word4	Word5
0	directory	working	files	repository	subversion
1	class	java	method	type	list
2	application	cocoa	mac	os	osx
3	data	error	function	list	type
4	best	code	error	vs	way
5	query	select	sql	list	xml
6	application	creator	widget	window	windows
7	bean	hibernate	mvc	security	framework
8	criteria	key	mapping	query	table
9	content	form	module	node	view
10	category	page	plugin	post	posts
11	database	pl	query	sql	table
12	add	custom	page	product	products
13	array	function	image	matrix	plot
14	asp	jquery	net	request	javascript
15	htaccess	mod_rewrite	php	rewrite	server
16	2007	custom	list	site	web
17	2005	2008	project	projects	solution
18	command	file	line	script	shell
19	cell	data	file	range	vba

將每群中的 title 記錄起來，分別去做 tf-idf，並將 max features dimension 限制為 5，如此一來，便可得到每群中較佳的 5 個字。

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

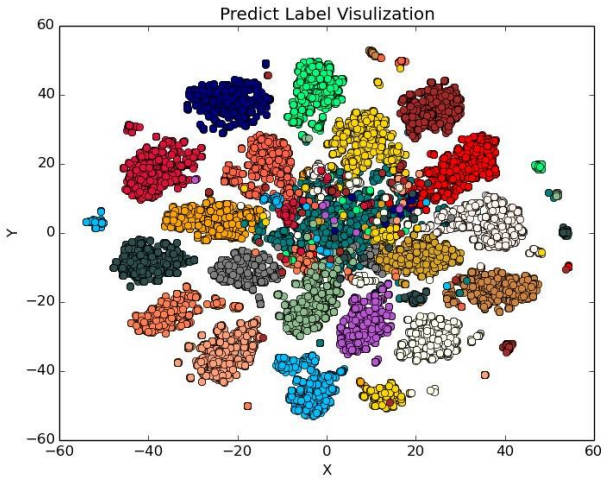


Fig1. Predict label

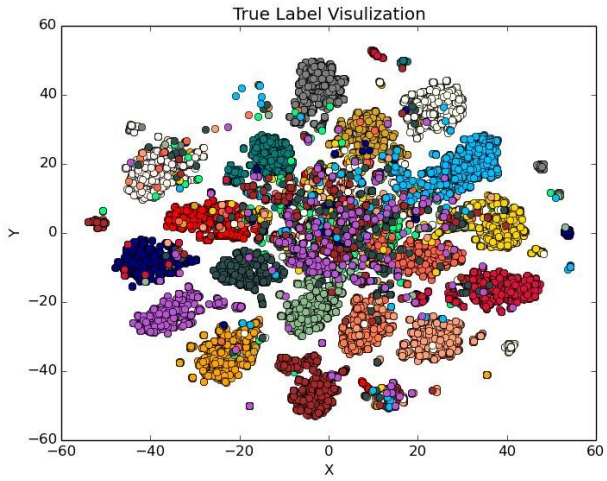
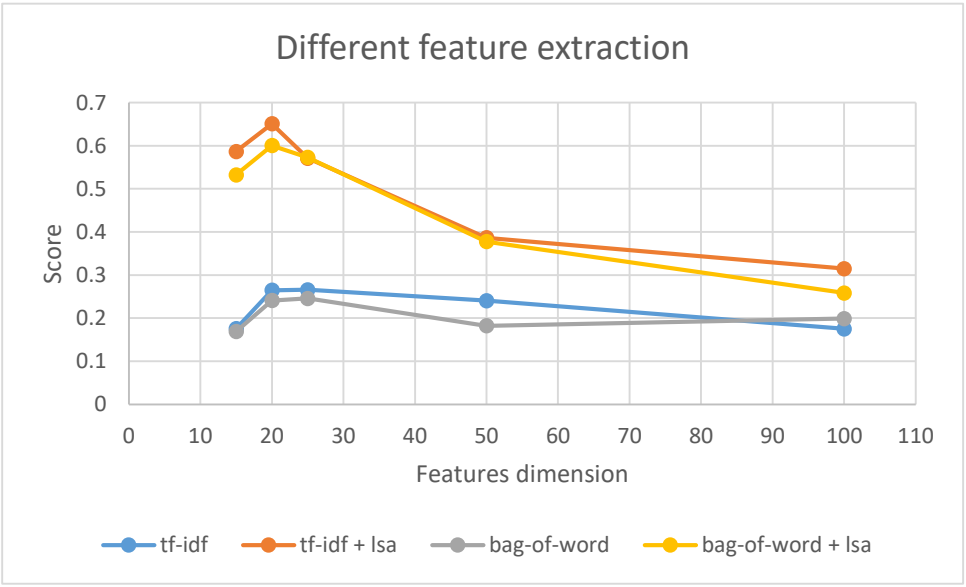


Fig2. True label

利用 Tf-idf + lsa 取出 20 維的 feature，再做 T-SNE 降到 2 維。

3. Compare different feature extraction methods.

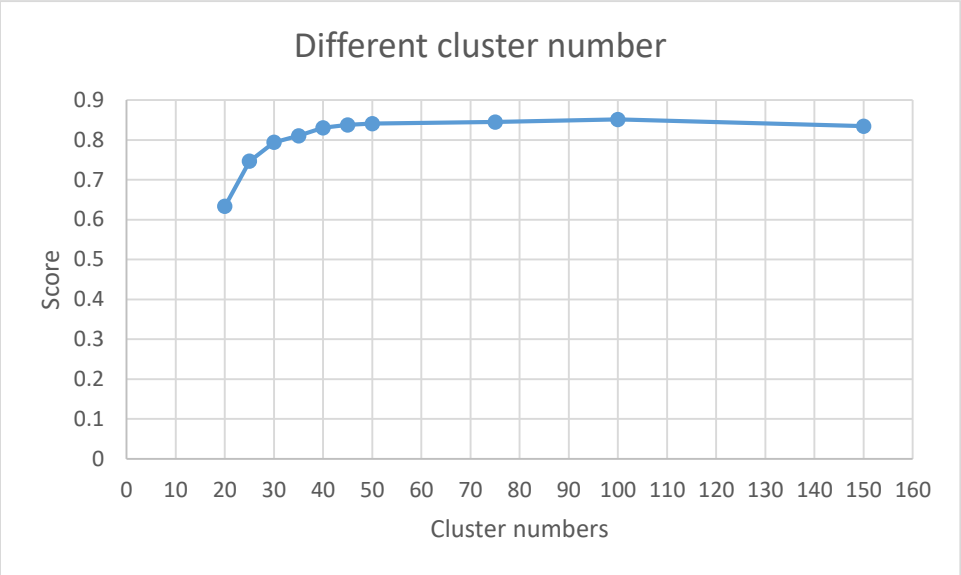


在本次作業中利用了 4 種不同的方法，觀察 kaggle private set score 的變化。縱坐標表示 kaggle private set score，橫坐標表示不同的 features dimension，分別為 15、20、25、50 及 100 維。Tf-idf 方法，直接限制了 maximum features dimension，而 Tf-idf + lsa 方法，則是做完 Tf-idf 後，做 lsa 降維到實驗的維度，Bag of word 與 Bag of word + lsa 亦同。

由結果來看，利用 lsa 降維會比一開始就限制 maximum features dimension 還要佳。而 Tf-idf 方

法普遍比 bag-of-word 還要佳。

4. **Try different cluster numbers and compare them.**



縱坐標為 kaggle private set score，橫坐標為 cluster numbers。本次作業中共有 20 種不同的 tag，利用 K-means 做分群，理想上，分為 20 群的結果會最佳，但實驗結果顯示，分到 100 群時，kaggle 的 score 會比只分 20 群時高了約 0.2，目前還無法想出為何會有這樣的結果。