Machine Learning 2016
homework 4 - Unsupervised Clustering & Dimensionality Reduction

翁丞世, R04945028, mob5566@gmail.com

Most Common Words in The Clusters

目標：分析給定之二十個群集中最常出現的字，對於每一群集列出出現頻率前十的字，列出的字皆以 stop words, bi-gram analysis (from gensim phrases model), stemming, tf-idf 篩選過.
附註：某些字會因為 stemming 而被縮短 (e.g. posts -> post, queries -> queri …)，而明顯與群集相關的字將會以紅色表示！

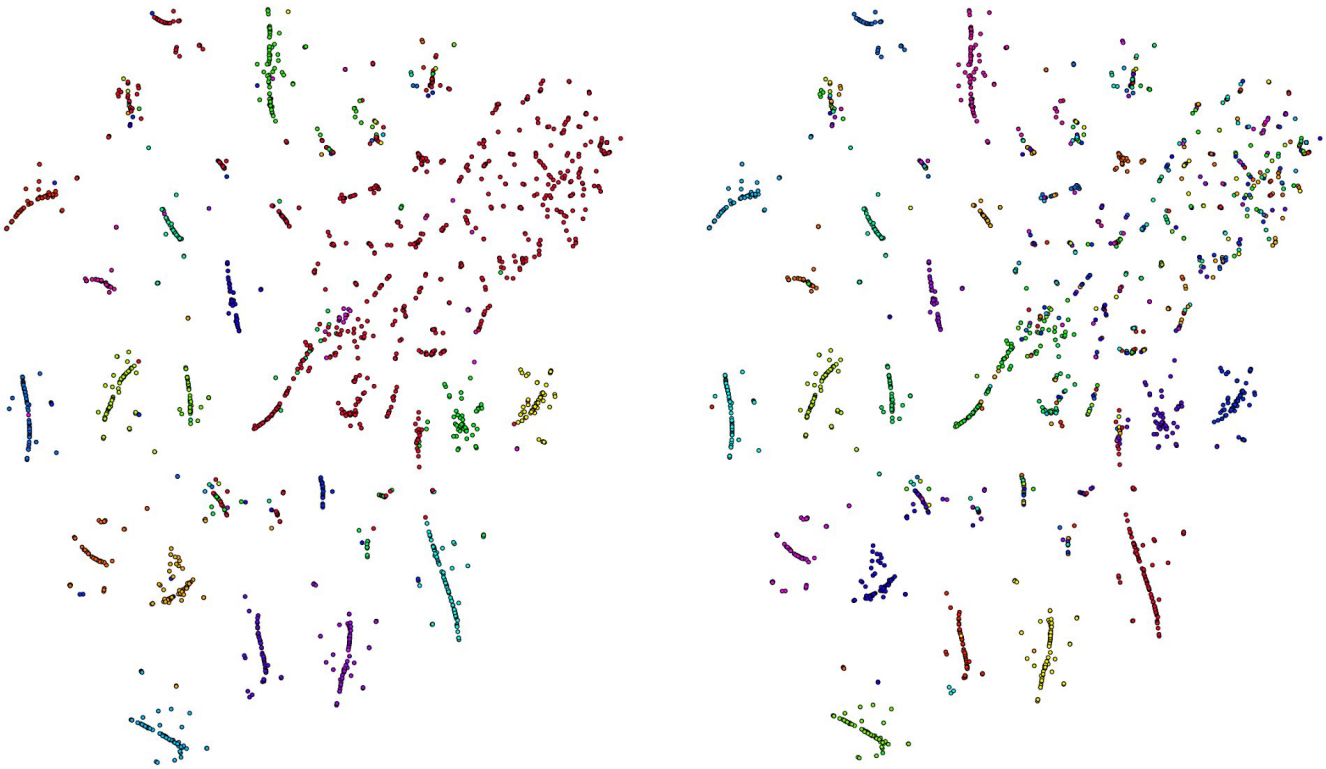
Cluster	Top 10 common words in each cluster (ordered by frequency)
wordpress	wordpress, post, word_press, page, categori, plugin, php, theme, wp, custom
oracle	oracl, sql, tabl, queri, connect, column, data, pl_sql, select, valu
svn	svn, subvers, commit, directori, svn_repositori, chang, repositori, server, revis, merg
apache	apach, mod_rewrit, redirect, server, php, url, rewrit, apach_mod, htaccess, request
excel	excel, cell, data, excel_vba, vba, row, rang, function, valu, column
matlab	matlab, function, matrix, imag, plot, arrai, vector, data, error, element
visual-studio	visual_studio, project, vs, solut, build, debug, add, code, window, visual
cocoa	cocoa, view, make, object, window, applic, creat, object_c, control, text
osx	mac, mac_os, os_x, osx, applic, window, instal, develop, mac_osx, app
bash	bash, bash_script, command, variabl, script, string, directori, line, output, run
spring	spring, bean, hibern, configur, spring_mvc, properti, spring_secur, annot, context, object
hibernate	hibern, map, queri, object, entiti, tabl, conllect, updat, problem, session
scala	scala, type, java, function, method, class, list, object, map, paramet
sharepoint	share_point, sharepoint, list, site, web_part, custom, creat, field, page, user
ajax	ajax, javascript, page, asp_net, load, php, ajax_request, jqueri, form, ajax_call
qt	qt, widget, window, applic, item, view, set, qt_creator, text, problem

drupal	drupal, node, view, form, page, modul, user, theme, custom, field
linq	linq, linq_sql, linq_queri, data, object, queri, list, group, select, valu
haskell	haskel, function, type, list, error, monad, problem, program, string, number
magento	magento, product, custom, categori, add, attribut, page, order, displai, chang

Visualization Clusters between My Result and Answer

方法描述：這邊的分析先將兩萬筆的標題依序以 python re (regular expression) 擷取出英文字母(去除標點符號)，接著由附件的 stop_words.txt 去除不相關的文字，再以 gensim PorterStemmer 把單字簡化 (e.g. posts -> post, queries -> queri…)，最後以 gensim phrases 將複合名詞結合起來 (e.g. visual studio -> visual_studio…)。處理好文字後，將兩萬筆標題再丟到 sklearn TfidfVectorizer 處理，將每行標題轉成 428 維的陣列，然後將這轉換過後的兩萬個陣列以 sklearn KMeans 分成二十群，此 model 能夠在 private set 上獲得 0.84 的分數。

視覺化：再把標題分類好之後，記錄其對應分類標籤．先將 428 維陣列以 sklearn NMF (Non-Negative Matrix Factorization) 降到 30 維，再隨機抽取 1000 個點以 sklearn tSNE 降到 2 維得以在二維平面上視覺化！視覺化結果如下，左圖為我的結果，右圖是正確標籤！



討論：自從使用 NMF 降維之後，可以明顯分出許多群，但是會發現，有一部分的標題以我的特徵截取方式並無法分開（左上分大塊紅色），應是缺乏明顯特徵單詞！如果能夠建立詞彙之間的相關性，再將此相關性套用到當前資料中，或許能夠分開.

Compare Different Feature Extraction Methods

描述：此部分討論方法皆都會先通過上一題之前處理來解取出不同維度之 tf-idf 陣列. 再將該陣列以不同方法降維.

方法	描述	F-score (private set)
pure tf-idf	如第二題描述	0.84
LSA	從 tf-idf 取得 3780 維的陣列，接著以 sklearn TruncatedSVD 將特徵降至 300 維，再以 k-means 分二十群，此方法稱為 Latent Semantic Analysis.	0.86
LDA	從 tf 中取得 3780 維的陣列，接著以 sklearn LatentDirichletAllocation 將維度降至 150 維，一樣以 k-means 分二十群，但此方法結果特別糟...，沒仔細去探究原因.	0.17
NMF	從 tf-idf 取得 3780 維的陣列，接著以 sklearn NMF 將特徵降至 25 維，NMF 會找到兩個非負的矩陣 (W, H) 相乘在一起會趨近於輸入的非負矩陣，再以 k-means 分二十群.	0.86

Different Cluster Numbers

描述：會第二題使用的模型來進行這一題的實驗，嘗試以不同的分群數來執行 k-means.

分析：根據第二題的視覺化結果，我認為應該增加分群數，能夠增加原本無法分開的部分的小分群，提升分辨率，有點類似某個大問題中，能夠再細分成許多小問題 (e.g. visual studio 中有問題是跟 debug 有關，有些則跟 solution 有關).

討論：下面由左而右分別是分成 20, 22, 25, 30 群，而 F-score 分別是 0.82, 0.85, 0.83, 0.79，發現稍微增加群集時，分數會有提升，但是增加太多反而又可能造成原本分好的群集被拆散，導致分數降低.