資料市集模組與開放資料貼標建模

Shunmao

2019/9/6

分析模組名稱

DataMarket

分析模組日期, 版本

Title: Automatically cleaning data and make a tag

Version: 1.0

Date: 2019-09-06

分析模組功能描述

這個分析模組主要是為了幫助國網中心資料市集平台(scidm)做csv檔的資料自動化清理，並且能分別好檔案與壞檔案。最後對資料作貼標籤的行為，目前標籤模型是用政府開放資料平台的標籤訓練而成。

套件導覽(參數與輸入輸出說明)

套件載入

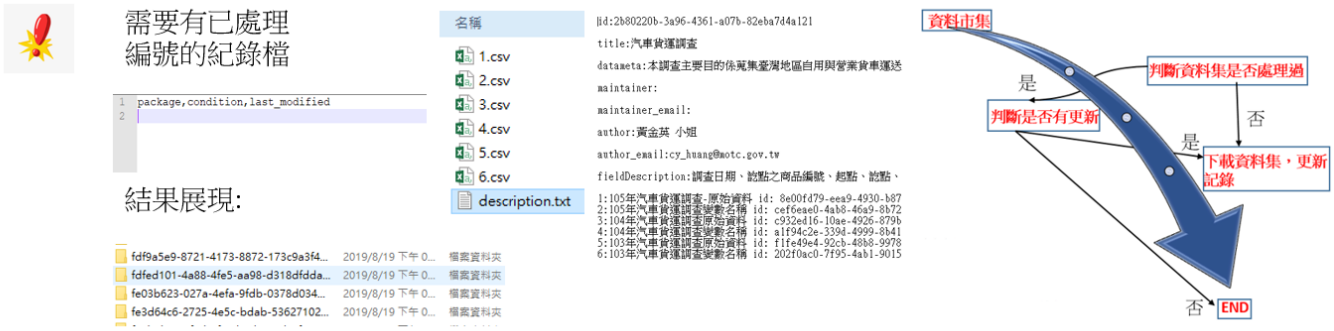
從scidm下載資料(Do_crawl)

資料市集平台 (https://scidm.nchc.org.tw/)本身採用ckan模組，運用R語言套件(ckanr)，做出自動資料爬取函數(Do_crawl)，參數如下圖所式，dealt.csv必須如圖所示，紀錄處理過資料集編號。需先load data.table,doParallel ckanr,lubridate等package，並且加載ckanr的網站。

ckanr_setup(url='https://scidm.nchc.org.tw/ (https://scidm.nchc.org.tw/)')

自動爬取程式(Do_crawl)

- Do_crawl(organ='政府開放資料',dealt_path=紀錄處理過的資料集編號,
  save_path=檔案存放的地點)



測試用的資料

測試資料為政府開放資料中的其中五筆，分別代表五種情況的代表資料:

第一種為刪除整行或整列為空值無意義變數或資料

第二種為檢測資料是否為HTML語法的檔案(開放資料中有很多誤存的檔案)

第三種為非csv格式檔案，卻存成csv檔案的壞資料格式，資料是用readLine讀取的 情況

第四種為重複變數名稱

第五種為使用錯誤編碼讀取檔案時的資料

```
data("test_data1");data("test_data2");data("test_data3")
data("test_data4");data("test_data5")
head(test_data1);head(test_data2);head(test_data3)
```

```
##            RptNo            RptName StatCourseNo
## 1: 10984-00-07-2 臺中市縱火案件分析  1098400a018
## 2: 10984-00-07-2 臺中市縱火案件分析  1098400a018
## 3: 10984-00-07-2 臺中市縱火案件分析  1098400a018
## 4: 10984-00-07-2 臺中市縱火案件分析  1098400a018
## 5: 10984-00-07-2 臺中市縱火案件分析  1098400a018
## 6: 10984-00-07-2 臺中市縱火案件分析  1098400a018
##                    StatCourseName            DataDate       PlaceNo
## 1: 縱火案件分析-縱火次數(單位:次) 2015-08-01T00:00:00 4.940689e-314
## 2: 縱火案件分析-縱火次數(單位:次) 2015-08-01T00:00:00 4.940689e-314
## 3: 縱火案件分析-縱火次數(單位:次) 2015-08-01T00:00:00 4.940689e-314
## 4: 縱火案件分析-縱火次數(單位:次) 2015-08-01T00:00:00 4.940689e-314
## 5: 縱火案件分析-縱火次數(單位:次) 2015-08-01T00:00:00 4.940689e-314
## 6: 縱火案件分析-縱火次數(單位:次) 2015-08-01T00:00:00 4.940689e-314
##    PlaceName PeriodNo PeriodName  Complex1      ComplexName  Complex2
## 1:    臺中市        M         月 600100001 第一救災救護大隊 603700001
## 2:    臺中市        M         月 600100001 第一救災救護大隊 603700002
## 3:    臺中市        M         月 600100001 第一救災救護大隊 603700003
## 4:    臺中市        M         月 600100001 第一救災救護大隊 603700004
## 5:    臺中市        M         月 600100001 第一救災救護大隊 603800001
## 6:    臺中市        M         月 600100001 第一救災救護大隊 603800002
##     Complex2Name Complex3 Complex3Name Complex4 Complex4Name Complex5
## 1:        建築物        0           NA        0           NA        0
## 2:          汽車        0           NA        0           NA        0
## 3:          機車        0           NA        0           NA        0
## 4: 縱火類別-其他        0           NA        0           NA        0
## 5:          明火        0           NA        0           NA        0
## 6:          汽油        0           NA        0           NA        0
##    Complex5Name DeriveNo DeriveExplain FValue SValue  RptDeptNo
## 1:           NA        0            NA      0     NA 387320000A
## 2:           NA        0            NA      0     NA 387320000A
## 3:           NA        0            NA      0     NA 387320000A
## 4:           NA        0            NA      0     NA 387320000A
## 5:           NA        0            NA      0     NA 387320000A
## 6:           NA        0            NA      0     NA 387320000A
##    RptDeptName              CreateTime ModifyTime
## 1:      消防局 2015-10-20T18:41:09.733         NA
## 2:      消防局 2015-10-20T18:41:09.737         NA
## 3:      消防局  2015-10-20T18:41:09.74         NA
## 4:      消防局  2015-10-20T18:41:09.74         NA
## 5:      消防局 2015-10-20T18:41:09.747         NA
## 6:      消防局 2015-10-20T18:41:09.747         NA
```

```
## [1] "<!DOCTYPE html>"
## [2] "<!--[if IE 7]> <html lang=\"zh_TW\" class=\"ie ie7\"> <![endif]-->"
## [3] "<!--[if IE 8]> <html lang=\"zh_TW\" class=\"ie ie8\"> <![endif]-->"
## [4] "<!--[if IE 9]> <html lang=\"zh_TW\" class=\"ie9\"> <![endif]-->"
## [5] "<!--[if gt IE 8]><!--> <html lang=\"zh_TW\"> <!--<![endif]-->"
## [6] "  <head>"
```

```
## [1] "PK\003\004\024"
## [2] "r\xcc\026g9`\xc3□\020'8v\xcf\030\xa3\001z\xd57\025~\024絜\xc8\x9a□\xfd(?湄\x91\027
\t:\035□\xe5\f)f□\"-A\031\b焍M\x921\xad\x9a□□\xaa\x81□疽6□蔪\001\037蝹5勺p襭\xf8\x98□插□\x9
9.~t烷□x\034\xcb\016胸□e\xee\x9d!k□e\024'R\017\xe4□H\xae\vZ遵$炟$晢\x9a9,@\xac\177q璗r燐\xca
\002H\xe2<虐□+kg\037J*□B\xe3□莸T\xd5\x92□□y朴在□e\036\006'P嬨\025\xf9\030F}堔\002Zy□.□P"
```

```
head(test_data4);head(test_data5)
```

```
##    期間 CN大陸(含香港)進口量(公噸) 其他國家進口量(公噸) 合計進口量(公噸)
## 1:  1月                  2854.8047             594.6753            82.8%
## 2:  2月                   983.8526             250.6635            79.7%
## 3:  3月                  2583.2810             253.9161            91.1%
## 4:  4月                  2602.0703             243.6541            91.4%
## 5:  5月                  2084.7439             264.4293            88.7%
## 6:  6月                  1997.6260             927.3310            68.3%
##    合計進口量(百分比) 合計進口量(公噸) 合計進口量(百分比)
## 1:              17.2%         3449.480                100%
## 2:              20.3%         1234.516                100%
## 3:               8.9%         2837.197                100%
## 4:               8.6%         2845.724                100%
## 5:              11.3%         2349.173                100%
## 6:              31.7%         2924.957                100%
```

```
##                              V1
## 1:             瑼□\xa1□\x90□迂
## 2: 103撟游漲瘜箇\xae□□敶□□.pdf
##                                                                  V2
## 1:                                                  銖□\xbc□\u0080\xa3蝯\x90
## 2: http://data.coa.gov.tw/GetOpenDataFile.aspx?id=964&FileType=DataMore&RID=150
```

函數SEARCH_ENCODE

這個函數主要是依據readr包的guess_encoding作為檔案之編碼判斷依據，選擇最高且機率大於20%的編碼，如果全部可能性皆小於20%，回傳unknown。輸入參數為檔案位置。

```
tmp =tempfile(fileext = '.csv')
write.csv(test_data1,tmp,row.names=F,fileEncoding = 'Big5')
print(SEARCH_ENCODE(tmp))
```

```
## [1] "Big5"
```

```
write.csv(test_data1,tmp,row.names=F,fileEncoding = 'UTF-8')
print(SEARCH_ENCODE(tmp))
```

```
## [1] "UTF-8"
```

```
unlink(tmp)
```

透過此函數，可知讀取未知檔案時，用什麼編碼讀取最合適。

Discri_HTML_File

此函數用來判斷錯誤存取的檔案，儲存格式為網頁程式碼。參數為檔案位置

```
tmp1=tempfile(fileext = '.csv')
tmp2=tempfile(fileext = '.csv')
writeLines(con = tmp1,text = test_data2)
write.csv(x = test_data1,file = tmp2,row.names = F)
Discri_HTML_File(tmp1)
```

```
## [1] TRUE
```

```
Discri_HTML_File(tmp2)
```

```
## [1] FALSE
```

```
unlink(tmp1);unlink(tmp2)
```

回傳值為邏輯True or False，代表此檔案是否為網頁程式碼誤存檔案。

Discri_worse_file

此函數是用來判斷儲存格式非csv檔(可能為pdf,word轉換成csv檔)，也就是無法讀取的檔案。 註:判別方法採用常用分隔符號(,和)，前10行分隔符號必須超過5個，才是正常的csv格式檔，需要注意的是把極小的好檔案誤判的風險。

```
tmp1=tempfile(fileext = '.csv')
tmp2=tempfile(fileext = '.csv')
writeLines(con = tmp1,text = test_data3)
write.csv(x = test_data1,file = tmp2,row.names = F)
Discri_worse_file(tmp1)
```

```
## [1] TRUE
```

```
Discri_worse_file(tmp2)
```

```
## [1] FALSE
```

```
unlink(tmp1);unlink(tmp2)
```

回傳邏輯值為True與False，判斷此檔案是否為無法讀取的檔案。

deter

此函數用data.table的fread函數作為判斷依據，如若讀取沒產生任何錯誤或警告即為good，警告為warning，而錯誤為error，將數據分成三種情況討論相對應的情形。

```
tmp1=tempfile(fileext = '.csv')
tmp2=tempfile(fileext = '.csv')
writeLines(con = tmp1,text = test_data2)
write.csv(x = test_data1,file = tmp2,row.names = F)
deter(tmp1)
```

```
## [1] "warning"
```

```
deter(tmp2)
```

```
## [1] "good"
```

```
unlink(tmp1);unlink(tmp2)
```

solve_problem_1

此函數功用為解決資料擁有重複變數名稱，會依序對重複的變數增加.1,.2,.3， 參數為dataframe格式

```
library(dplyr)
colnames(test_data4)
```

```
## [1] "期間"                          "CN大陸(含香港)進口量(公噸)"
## [3] "其他國家進口量(公噸)"          "合計進口量(公噸)"
## [5] "合計進口量(百分比)"            "合計進口量(公噸)"
## [7] "合計進口量(百分比)"
```

```
solve_problem_1(test_data4) %>% colnames
```

```
## [1] "期間"                          "CN大陸(含香港)進口量(公噸)"
## [3] "其他國家進口量(公噸)"          "合計進口量(公噸)"
## [5] "合計進口量(百分比)"            "合計進口量(公噸).1"
## [7] "合計進口量(百分比).1"
```

rm_rowANDcolumn_bad

此函數功用為去除無意義變數或資料，也就是全為空值("",NA)

```
test_data1 %>% dim
```

```
## [1] 207  27
```

```
apply(test_data1,2,is.na) %>% apply(2,sum)
```

```
##          RptNo        RptName   StatCourseNo StatCourseName       DataDate
##              0              0              0              0              0
##        PlaceNo      PlaceName       PeriodNo     PeriodName       Complex1
##              0              0              0              0              0
##    ComplexName       Complex2   Complex2Name       Complex3   Complex3Name
##              0              0              0              0            207
##       Complex4   Complex4Name       Complex5   Complex5Name       DeriveNo
##              0            207              0            207              0
##  DeriveExplain         FValue         SValue      RptDeptNo    RptDeptName
##            207              0            207              0              0
##     CreateTime     ModifyTime
##              0            207
```

```
rm_rowANDcolumn_bad(test_data1) %>% dim
```

```
## [1] 207  21
```

此資料有6個變數(Complex3Name,Complex4Name,Complex5Name, DeriveExplain,SValue,ModifyTime)皆是整行空值，被此函數刪除。

Test_encode

此函數功用為測試編碼是否正確讀取，如未正確讀取，R內建base的 nchar(type='nchars')可能會產生錯誤，因此可用作為偵錯的依據。

```
#good encoding
print(TEST_ENCODE(test_data1))
```

```
## [1] "OK"
```

```
#wrong encoding
print(TEST_ENCODE(test_data5))
```

```
## [1] "Maybe wrong encode"
```

Discriminant

測試資料是否有至少有一個觀測值且2個變數

Auto_clean(Auto_clean2)

這個包裡面最重要的函數之一，也是上述功能集合的函數。除了上述功能以外，會刪除公開資料特有現象，把資料描述寫進csv檔裡，另外如果deter為warning或 error時，會採用補救函數Auto_clean2，首先猜測資料的分隔符號是什麼，並依據資料中每一行擁有分隔符號數量最多的比例作為刪除依據，刪除分隔符號量與之不同的行，使檔案可讀取(會遺失資訊)。參數為檔案位置

Pred_fun

這個包裡面重要函數之一，用描述文字預測此資料集需要貼什麼標籤，透過結巴加載大量詞庫做文字斷詞，做成dtm矩陣，並建立分類預測模型，共採用三個模型，naive bayes,隨機森林和svm with linear kernel，參數為文字，回傳為 15個類別標籤。詳細過程可參考附錄。

```
seg=act_seg()
Pred_fun('澎湖縣觀光遊憩人次資料')
```

```
## [1] "休閒旅遊"
```

```
Pred_fun('長期照顧管理中心')
```

```
## [1] "退休安養"
```

附錄

Data

1.針對政府開放資料，最後下載了16577筆的Title(資料集名稱)+datameta(資料集描述)，其中有681筆資料集名稱重複，透過隨機選取刪除重複標題，剩餘15896筆。

2.各國政府開放資料，採用分類相似，採用政府開放資料平台上的原始分類，透過額外的爬蟲方式，搜尋資料集的類別，其中有942筆資料集已刪除，無法判別，剩餘14954筆。

3.有2筆Title，詞為英文和停止詞組成，刪除，剩餘14952筆 (嘉義縣政府itaiwan、Financial industry consolidation M&A Deals since September 2004)

4.有23筆Title，刪除停止詞後，剩餘單字皆為長度1的單字，剩餘14929筆

資料處理

使用R語言套件中的jiebaR，對文字做斷詞。

新增各式停止詞: 常見用字:一樣,一般,一轉眼,萬一,上,上下,下,不,不僅,不但,不光,不單,不只,不外乎,…

數字:0,1,2,3,4,…

英文:a-zA-Z

標點符號:，,、,?,。,$,…

台灣地名:花蓮,台中,臺中,宜蘭,彰化,萬里,金山,板橋,汐止,深坑,石碇,瑞芳,平溪,…

時間(副詞):年,月,日,時,分,秒,最近,曾經,過去,現在,未來,…

常見常用但無幫助的字:資料表,數據,資料,政府,本署,報表,…

無意義的字:別分,天內,數按,…

新增各式字彙使斷詞更為精準:

新增台灣專屬詞庫，約24萬筆單字。

新增自建辭彙，使斷詞更為精確。

新增部分搜狗詞庫

重組資料

投資理財	419
公共資訊	400
生活安全及品質	400
交通及通訊	395
求學及進修	340
就醫	289
求職及就業	243
購屋及遷徙	193
休閒旅遊	184
生育保健	124
服兵役	90
退休安養	75
開創事業	60
生命禮儀	46
選舉及投票	16

因資料類別不均勻，與部分類別標錯，導致分類困難，採取重組資料，讓資料品質提升。

將出生及收養+婚姻+生命禮儀定義為生命禮儀，老人安養和退休合併為退休安養，類別從18項變成15項 。人工盡可能找尋每一個類別的關鍵字，找出不超過400筆的代表性數據。

(找尋的關鍵字可參閱附錄)

修正 開創事業包含一些標題分類錯誤，做修正後，投資理財多增加了19筆數據，超過400筆。

建模過程

切割資料集，80%為訓練集，剩餘20%為測試集。

訓練集:2444筆 測試集:628筆

將訓練集做結巴斷詞，共6287個單字變數

運用全變數+naivebayes (不展現confusion matrix，15*15過於凌亂)

```
Overall Statistics

               Accuracy : 0.8658
                 95% CI : (0.8366, 0.8915)
    No Information Rate : 0.1294
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8516
```

Train accuracy:98.81% Test accuracy:86.58%

運用全變數+randomForest (不展現confusion matrix，15*15過於凌亂)

```
Overall Statistics

               Accuracy : 0.8706
                 95% CI : (0.8418, 0.8959)
    No Information Rate : 0.1294
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8566
```

Train accuracy:96.64% Test accuracy:87.06%

運用全變數svm (linear kernel) (不展現confusion matrix，15*15過於凌亂)

```
               Accuracy : 0.901
                 95% CI : (0.8748, 0.9232)
    No Information Rate : 0.1294
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8902
```

Train accuracy:99.71% Test accuracy:90.1%

NB

Statistics by Class:

	Class: 公共資訊	Class: 生育保健	Class: 生命禮儀	Class: 生活安全及品質	Class: 交通及通訊
Sensitivity	0.67470	0.85714	0.70000	0.9324	0.88525
Specificity	0.96881	0.99012	1.00000	0.9639	0.98236
Pos Pred Value	0.76712	0.75000	1.00000	0.7753	0.84375
Neg Pred Value	0.95135	0.99503	0.99517	0.9907	0.98759
Prevalence	0.13217	0.03344	0.01592	0.1178	0.09713
Detection Rate	0.08917	0.02866	0.01115	0.1099	0.08599
Detection Prevalence	0.11624	0.03822	0.01115	0.1417	0.10191
Balanced Accuracy	0.82175	0.92363	0.85000	0.9482	0.93380

	Class: 休閒旅遊	Class: 投資理財	Class: 求學及進修	Class: 求職及就業	Class: 服兵役
Sensitivity	0.84615	0.9647	0.9714	0.80357	0.91667
Specificity	0.98472	0.9890	0.9821	0.98776	1.00000
Pos Pred Value	0.78571	0.9318	0.8718	0.86538	1.00000
Neg Pred Value	0.98976	0.9944	0.9964	0.98090	0.99838
Prevalence	0.06210	0.1354	0.1115	0.08917	0.01911
Detection Rate	0.05255	0.1306	0.1083	0.07166	0.01752
Detection Prevalence	0.06688	0.1401	0.1242	0.08280	0.01752
Balanced Accuracy	0.91544	0.9768	0.9768	0.89567	0.95833

	Class: 退休安養	Class: 就醫	Class: 開創事業	Class: 選舉及投票	Class: 購屋及遷徙
Sensitivity	0.58824	0.81633	0.89474	1.000000	0.89286
Specificity	0.99509	1.00000	0.99836	1.000000	1.00000
Pos Pred Value	0.76923	1.00000	0.94444	1.000000	1.00000
Neg Pred Value	0.98862	0.98469	0.99672	1.000000	0.99502
Prevalence	0.02707	0.07803	0.03025	0.006369	0.04459
Detection Rate	0.01592	0.06369	0.02707	0.006369	0.03981
Detection Prevalence	0.02070	0.06369	0.02866	0.006369	0.03981
Balanced Accuracy	0.79166	0.90816	0.94655	1.000000	0.94643

RF

Statistics by Class:

	Class: 公共資訊	Class: 生育保健	Class: 生命禮儀	Class: 生活安全及品質	Class: 交通及通訊
Sensitivity	0.9157	0.90476	0.80000	0.8919	0.95082
Specificity	0.9505	1.00000	0.99838	0.9964	0.99471
Pos Pred Value	0.7379	1.00000	0.88889	0.9706	0.95082
Neg Pred Value	0.9867	0.99672	0.99677	0.9857	0.99471
Prevalence	0.1322	0.03344	0.01592	0.1178	0.09713
Detection Rate	0.1210	0.03025	0.01274	0.1051	0.09236
Detection Prevalence	0.1640	0.03025	0.01433	0.1083	0.09713
Balanced Accuracy	0.9331	0.95238	0.89919	0.9441	0.97276

	Class: 休閒旅遊	Class: 投資理財	Class: 求學及進修	Class: 求職及就業	Class: 服兵役
Sensitivity	0.84615	0.9294	0.81429	0.82143	1.00000
Specificity	0.99660	0.9908	0.98746	0.98776	1.00000
Pos Pred Value	0.94286	0.9405	0.89063	0.86792	1.00000
Neg Pred Value	0.98988	0.9890	0.97695	0.98261	1.00000
Prevalence	0.06210	0.1354	0.11146	0.08917	0.01911
Detection Rate	0.05255	0.1258	0.09076	0.07325	0.01911
Detection Prevalence	0.05573	0.1338	0.10191	0.08439	0.01911
Balanced Accuracy	0.92138	0.9601	0.90087	0.90460	1.00000

	Class: 退休安養	Class: 就醫	Class: 開創事業	Class: 選舉及投票	Class: 購屋及遷徙
Sensitivity	0.76471	0.95918	0.84211	1.000000	1.00000
Specificity	0.99673	0.99655	0.99015	1.000000	0.99667
Pos Pred Value	0.86667	0.95918	0.72727	1.000000	0.93333
Neg Pred Value	0.99347	0.99655	0.99505	1.000000	1.00000
Prevalence	0.02707	0.07803	0.03025	0.006369	0.04459
Detection Rate	0.02070	0.07484	0.02548	0.006369	0.04459
Detection Prevalence	0.02389	0.07803	0.03503	0.006369	0.04777
Balanced Accuracy	0.88072	0.97786	0.91613	1.000000	0.99833

SVM

Statistics by Class:

	Class: 公共資訊	Class: 生育保健	Class: 生命禮儀	Class: 生活安全及品質	Class: 交通及通訊
Sensitivity	0.9375	0.79167	0.500000	0.9259	0.89062
Specificity	0.9359	0.99336	0.998377	0.9908	0.99644
Pos Pred Value	0.6818	0.82609	0.833333	0.9375	0.96610
Neg Pred Value	0.9903	0.99171	0.991935	0.9890	0.98765
Prevalence	0.1278	0.03834	0.015974	0.1294	0.10224
Detection Rate	0.1198	0.03035	0.007987	0.1198	0.09105
Detection Prevalence	0.1757	0.03674	0.009585	0.1278	0.09425
Balanced Accuracy	0.9367	0.89251	0.749188	0.9584	0.94353

	Class: 休閒旅遊	Class: 投資理財	Class: 求學及進修	Class: 求職及就業	Class: 服兵役
Sensitivity	0.80556	0.9875	0.85507	0.91667	0.93750
Specificity	1.00000	0.9982	0.98923	0.98962	1.00000
Pos Pred Value	1.00000	0.9875	0.90769	0.88000	1.00000
Neg Pred Value	0.98827	0.9982	0.98217	0.99306	0.99836
Prevalence	0.05751	0.1278	0.11022	0.07668	0.02556
Detection Rate	0.04633	0.1262	0.09425	0.07029	0.02396
Detection Prevalence	0.04633	0.1278	0.10383	0.07987	0.02396
Balanced Accuracy	0.90278	0.9928	0.92215	0.95314	0.96875

	Class: 退休安養	Class: 就醫	Class: 開創事業	Class: 選舉及投票	Class: 購屋及遷徙
Sensitivity	0.68750	0.96491	0.81250	1.00000	0.96000
Specificity	0.99836	0.99824	1.00000	1.00000	1.00000
Pos Pred Value	0.91667	0.98214	1.00000	1.00000	1.00000
Neg Pred Value	0.99186	0.99649	0.99511	1.00000	0.99834
Prevalence	0.02556	0.09105	0.02556	0.00639	0.03994
Detection Rate	0.01757	0.08786	0.02077	0.00639	0.03834
Detection Prevalence	0.01917	0.08946	0.02077	0.00639	0.03834
Balanced Accuracy	0.84293	0.98158	0.90625	1.00000	0.98000

模型選擇

最終模型選擇:

透過各種統計指標(Recall,Precision,Accuracy,...)SVM表現普遍較佳，因此選SVM為主要預測模型。然而SVM在'生命禮儀'這項類別中，平均表現不是這麼好，因此當其他兩個模型預測一致時，採用其他兩模型預測結果。(共292筆不一樣，目測兩模型一致時，貼的標籤較吻合)