

使用 NER 模型產生的風險詞，並加上風險詞的上下文，以句點作為結尾。將這個資料作為訓練集，訓練隨機森林分類模型

分別對每個案件類型建立分類模型

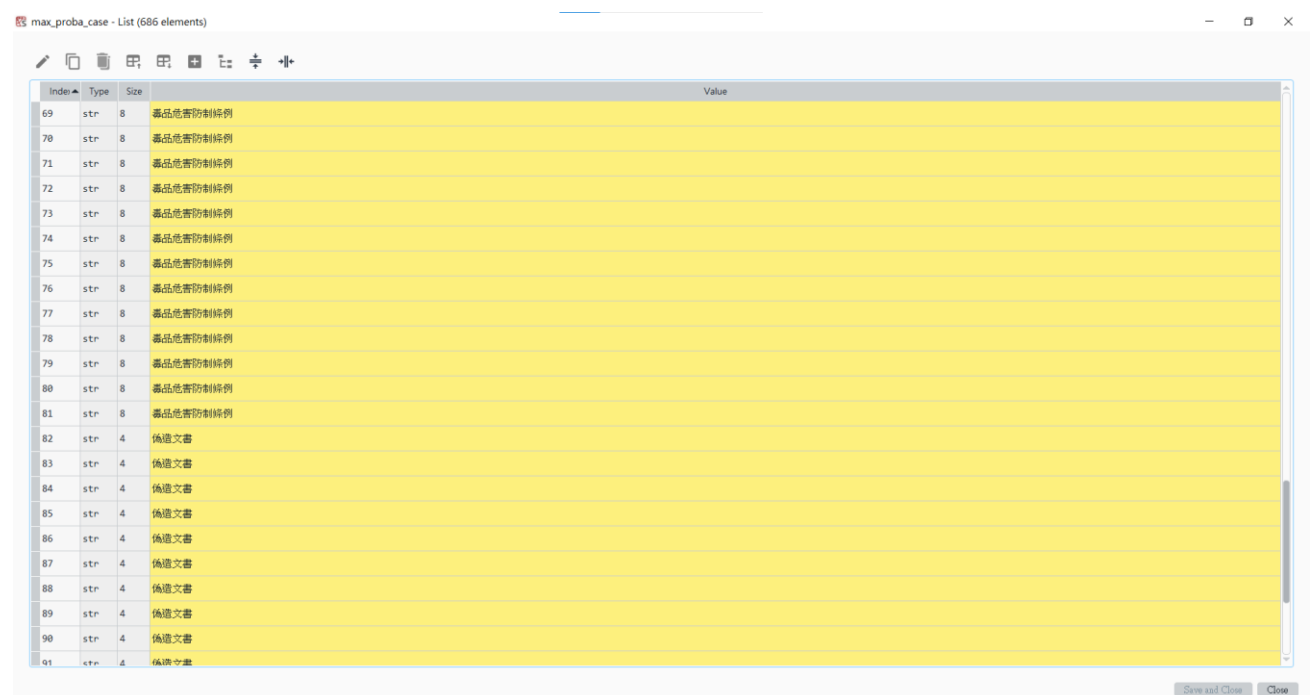
```
#Use Randomforest to train classifier model
rf_classifier_seg_drug = RandomForestClassifier()
rf_classifier_seg_forged_documents = RandomForestClassifier()
rf_classifier_seg_gamble = RandomForestClassifier()
rf_classifier_seg_negligent_injury = RandomForestClassifier()
rf_classifier_seg_public = RandomForestClassifier()
rf_classifier_seg_theft = RandomForestClassifier()

rf_classifier_seg_drug.fit(train_embedding, merged_df['label_case1']==1)
rf_classifier_seg_forged_documents.fit(train_embedding, merged_df['label_case2']==1)
rf_classifier_seg_gamble.fit(train_embedding, merged_df['label_case3']==1)
rf_classifier_seg_negligent_injury.fit(train_embedding, merged_df['label_case4']==1)
rf_classifier_seg_public.fit(train_embedding, merged_df['label_case5']==1)
rf_classifier_seg_theft.fit(train_embedding, merged_df['label_case6']==1)
```

將上述的資料轉換成詞嵌入向量，作為輸入屬性，而目標屬性分別為「是案件（1）/非案件（0）」，各別對應六種案件，例如：是竊盜（1）/非竊盜（0）。將測試資料輸入六個訓練好的分類模型，就會得到六組是非的機率值（predict\_proba），再取出最高「是」的機率轉換成對應的案件，以此作為分類預測結果

```
probs = case[j].predict_proba(test)
probs_all_list.append(probs)
```

預測結果如下：



Index	Type	Size	Value
69	str	8	毒品危害防制條例
70	str	8	毒品危害防制條例
71	str	8	毒品危害防制條例
72	str	8	毒品危害防制條例
73	str	8	毒品危害防制條例
74	str	8	毒品危害防制條例
75	str	8	毒品危害防制條例
76	str	8	毒品危害防制條例
77	str	8	毒品危害防制條例
78	str	8	毒品危害防制條例
79	str	8	毒品危害防制條例
80	str	8	毒品危害防制條例
81	str	8	毒品危害防制條例
82	str	4	偽造文書
83	str	4	偽造文書
84	str	4	偽造文書
85	str	4	偽造文書
86	str	4	偽造文書
87	str	4	偽造文書
88	str	4	偽造文書
89	str	4	偽造文書
90	str	4	偽造文書
91	str	4	偽造文書

另外使用同一種方式，但不經過 NER 模型，直接將原文轉換成詞嵌入向量，訓

練隨機森林模型，將這兩個模型進行比較

上下文：

詞嵌入(分類)	Period
F1_micro	0.9723
F1_macro	0.9685

原文：

詞嵌入(分類)	Original
F1_micro	0.8921
F1_macro	0.8701

與分類模型相似，將包含上下文的風險詞資料轉換成詞嵌入向量作為輸入屬性，以刑期作為目標屬性建立的隨機森林回歸模型，及以罰金作為目標屬性的雖機森林回歸模型，最後使用 **MSE** 進行評估。

預測結果：

y_pred_fine - NumPy object array		y_pred_years - NumPy object array	
	0		0
0	0	0	0.1913
1	500	0	0.1599
2	0	0	0.1755
3	400	0	0.3322
4	0	0	0.2031
5	0	0	0.3206
6	0	0	0.1599
7	0	0	0.25
8	0	0	0.1971
9	180	0	0.2286
10	0	0	0.1571

評估結果：

Public_Danger	( 總篇數:1331 , 訓練:1065 , 測試:266 )					Forged_Documents	( 總篇數:406 , 訓練:325 , 測試:81 )				
max_features	3000	4000	5000	6000	7000	max_features	3000	4000	5000	6000	7000
MSE	0.0324	0.0292	0.0338	0.0349	0.0282	MSE	0.2283	0.2061	0.2121	0.2037	0.2096
Drug	( 總篇數:411 , 訓練:329 , 測試:82 )					Gamble	( 總篇數:408 , 訓練:326 , 測試:82 )				
max_features	3000	4000	5000	6000	7000	max_features	3000	4000	5000	6000	7000
MSE	5.3521	4.9716	5.1418	5.1905	5.1317	MSE	0.0399	0.0391	0.04	0.0403	0.0392
Theft	( 總篇數:409 , 訓練:327 , 測試:82 )					Negligent_Injury	( 總篇數:467 , 訓練:374 , 測試:93 )				
max_features	3000	4000	5000	6000	7000	max_features	3000	4000	5000	6000	7000
MSE	0.2274	0.2062	0.218	0.2183	0.222	MSE	0.006	0.0059	0.0066	0.0065	0.0063