創新變革精進獎勵項目

# 資料蒐集之改造利器－自動判讀檢察書類智慧模組

檢察機關使用大量法律書類，各項犯罪統計多由書類文字蒐集而來，統計人員逐案翻卷建置資料，已難以應付快速擴增之業務量能，為提升工作效率，爰運用專業創建「自動判讀檢察書類智慧模組」，並建立「情境資料庫」，以提升檢察統計效能及增進資料應用彈性。

**法務部統計處**（臺灣高等檢察署統計室林科長在一、余科長瑞珺）

## 壹、前言

檢察機關統計人員負責蒐集犯罪資料，各類刑事案件辦理終結後，書記官會將相關書類（包含起訴書、法院裁判書……等）隨同卷宗交由統計人員逐案翻卷蒐集資料，隨著刑事案件量大幅成長，所需之犯罪統計愈趨多樣，工作負荷日益加重。

鑑於法務部致力科技化，並將「AI 輔助提升檢察效能」訂為施政目標，爰發想運用自然語言處理（Natural Language Processing, NLP），建置「自動判讀檢察書類智慧模組」，改造資料蒐集流程，並建立「情境資料庫」，增進資料應用彈性。

## 貳、運用科學方法，改造資料蒐集流程，提升檢察統計效能

NLP 是人工智慧（AI）領域之一，使用機器學習技術來處理及解讀文字和資料，適用於檢察機關案件書類，法務統計人員乃運用 NLP 創建智慧模組，自動判讀書類，並將判讀結果匯入系統，減省資料蒐集時間；另一方面，將模組結合公務統計系統，建立「情境資料庫」，以即時支援業務單位回應民眾關切議題。（下頁圖 1）

### 一、創建「自動判讀檢察書類智慧模組」

（一）斷詞提取關鍵字

將檢察書類內容（案號、被告、終結情形、罪名、

法條、犯罪手法等）以CKIP[1]進行斷詞，拆分成單一詞彙，再以 python 程式語言排除重複詞彙後，提取關鍵字作爲初步解釋變數，並將其轉換爲布林值（1 或 0）。

（二）篩選變數

利用 SQL 程式語言查詢資料庫中之實際資料，作爲目標變數，再運用屬性子集評估器（CFS）[2]自初步解釋變數中，篩選出最終解釋變數。

（三）建立及驗證模型

運用資料探勘軟體 WEKA[3]進行建模實驗，採用決策樹、隨機森林、KNN（K- 近鄰演算法）、羅吉斯迴歸與集成分析 5 種方法建立統計模型，選定最適模型爲「決策樹演算法」，正確率達 81.943％，使用 R 語言繪製 ROC 曲線[4]，模型判斷價值爲中上。（下頁圖 2）

圖 1　「自動判讀檢察書類智慧模組」建置及運用


資料來源：作者自行繪製。

## 二、改造資料蒐集流程

透過「自動判讀檢察書類智慧模組」，判讀檢察書類文字，轉化爲系統欄位資料，匯入刑案資訊整合系統，進行資料蒐集；與業務單位收案資料勾稽校正；再利用公務統計系統之檢核功能，依過往資料結構及合理性，篩選出邏輯錯誤之案件，以確保資料正確性。（第 97 頁圖 3）

## 三、結合公務統計系統，建立「情境資料庫」

實務上突發之重大輿情，若爲未常川蒐集之案類，無法即時提供業務單位予以回應，爲解決此問題，爰運用智慧模組之功能，彈性增加書類關鍵字之決策樹模型，判別案件類別後，於公務統計系統建立對應之「情境資料表」，並整合爲「情境資料庫」，隨時產製統計數據即時應用。

# 創新變革精進獎勵項目

## 參、成果效益

### 一、自動判讀書類，減省資料蒐集時間，提升工作效能

檢察書類透過「自動判讀檢察書類智慧模組」，可自動產生判讀結果，匯入刑案資訊整合系統相對應之欄位，並進行校正及檢誤，平均每件輸入時間縮短至 5 分鐘以內。

圖 2　「自動判讀檢察書類智慧模組」建置流程


資料來源：作者自行繪製。

### 二、模組解讀檢察書類標準一致，降低人為誤差，強化資料品質

「自動判讀檢察書類智慧模組」具相同之解讀標準，判讀結果不受案件複雜度、書類用語差異之影響，錯誤率由 9.4%降低至 3.5%。

### 三、多元「情境資料庫」，即刻支援業務，解燃眉之急

未常川蒐集之案類，可依業務單位提供之關鍵字，從檢察書類查詢系統篩選出相關書類，於「自動判讀檢察書類智慧模組」增加決策樹模型，建置對應之「情境資料表」，再整合至「情境資料庫」，可於 2 天內產製數據，即刻支援業務。

## 肆、革新性

### 一、首度運用自然語言處理技術判讀書類，呼應法務數位轉型

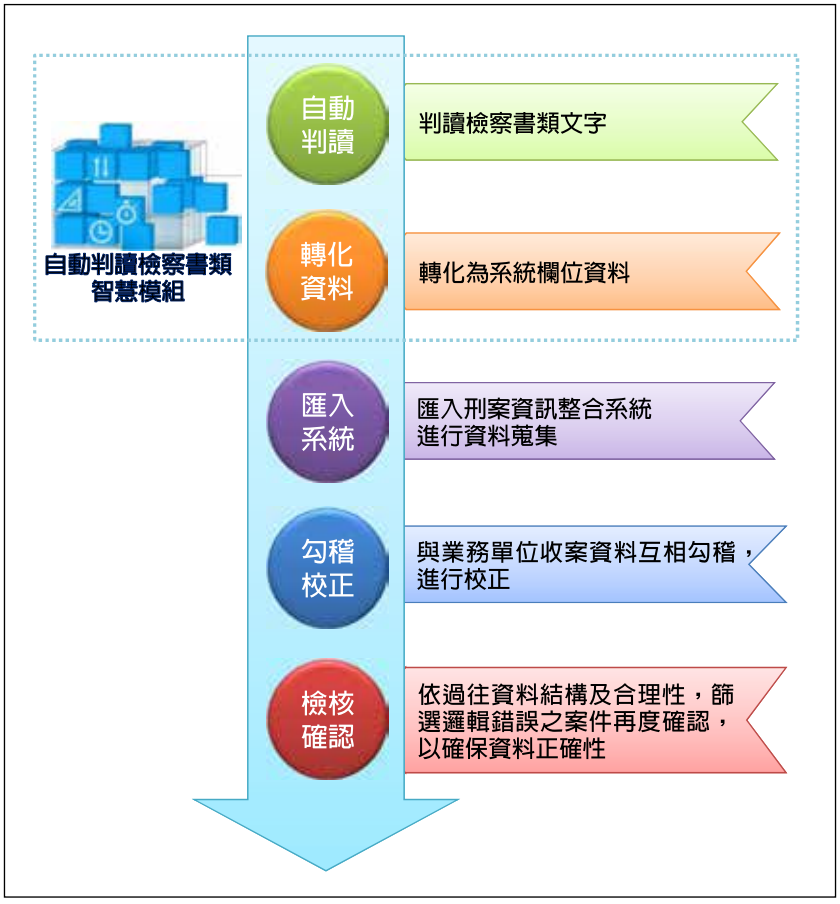
首度運用自然語言處理技術，訓練人工智慧模型判讀書

類文本，擷取所需之特定資訊，匯入刑案資訊整合系統，並利用相關資料庫系統進行校正與檢核，改造資料蒐集流程，取代傳統作業模式，呼應法務數位轉型。

## 二、跳脫傳統資料蒐集框架，創造統計與業務雙贏

依業務單位提供之關鍵字，篩選檢察書類，再利用模組彈性增加決策樹模型，同步擴增「情境資料庫」之類別，不受系統欄位增設程序之限制，迅速產製統計數據，即時支援業務單位回應民眾關切議題，創造統計與業務雙贏。

圖 3　改造資料蒐集流程




資料來源：作者自行繪製。

## 伍、結語

法務統計人員發揮專業能力，活用多種數據分析工具與方法，打造智慧模組，改造資料蒐集流程，減輕工作負擔，並擴增智慧模組之用途，迅速產製數據支援業務，有效提升檢察統計服務量能。

## 註釋

1. CKIP（Chinese Knowledge and Information Processing）為中央研究院詞庫小組開發之自然語言處理繁體中文斷詞工具。
2. 屬性子集評估器指 Correlation-based Feature Selection，是機器學習特徵選擇的方法，可篩選出與目標變數高度相關且彼此不相關之解釋變數。
3. WEKA 指 Waikato Environment for Knowledge Analysis，是紐西蘭懷卡托大學用 Java 開發的資料探勘常用軟體，包含完整的資料探勘處理流程。
4. ROC 指 Receiver operating characteristic curve)，曲線是一種坐標圖式的分析工具，可通過圖示觀察模型準確性。❖