

GAI Final Project: Written judgment summarization

C14094071 陳柏宏

F74092269 陳冠廷

F74092235 林晉德

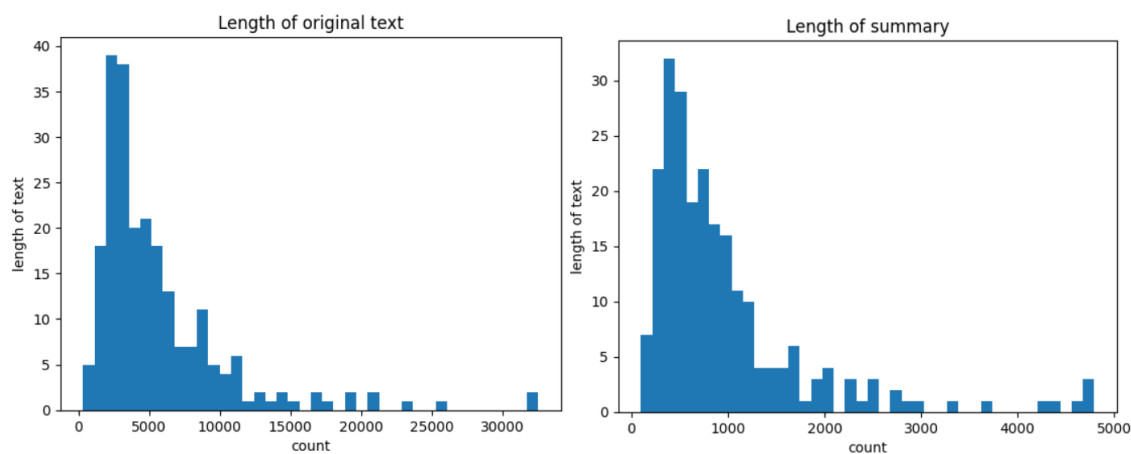
1. Data preprocessing

因為我們認為判決原文為資料中最重要的部分，其他的欄位如爭點、裁判案由.....等皆只是從裁判原文中提取出來的部分內容，不一定跟摘要要有絕對相關。所以以下的資料處理集中在分析、處理「判決原文」的部分。

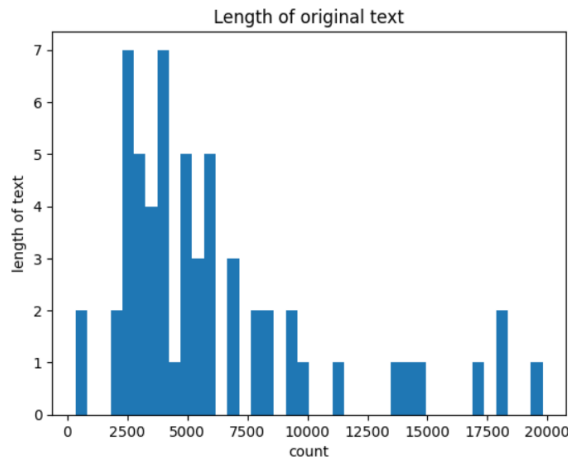
I、文字長度分析

分析後可以得知 Training data 資料有 236 筆。

裁判原文多集中在 5000 字以下，摘要多為 1500 字以下。



Testing data 則有 59 筆，裁判原文多集中在 6500 字以下。



II、移除雜訊

- i. 刪除其中 6 筆過短資料(noise)。

由以上的裁判原文的長度分析中可以得出，有 6 筆過短資料，其字數都低於 400 字。查看其原文後可以發現這 6 筆資料中都有寫到「連結以取得判決全文」(如下圖所示)，可以得知皆為無用資料，所以我們將其視為雜訊並移除。所佔比例為 $6/236=2.5\%$ ，影響不小。

連結以取得判決全文。中華民國106年4月13日最高行政法院第一庭審判長法官劉鑫楨法官胡方新法官程怡怡法官張國勳法官汪漢卿以上正本證明與原本無異中華民國106年4月13日由上訴人負擔。理由請點選下方附件連結以取得判決全文。中華民國105年11月24日最高行政法院第一庭審判長法官劉鑫楨法官胡方新法官程怡怡法官張國勳法官汪漢卿以上正本證明與原本無異中華民國105年11月24日由上訴人負擔。理由請點選下方附件連結以取得判決全文。中華民國106年6月8日最高行政法院第一庭審判長法官劉鑫楨法官胡方新法官程怡怡法官張國勳法官汪漢卿以上正本證明與原本無異中華民國106年6月8日由上訴人負擔。理由請點選下方附件連結以取得判決全文。中華民國105年7月14日智慧財產法院105年度行商訴字第18號行政判決。提起上訴。本院判決如下：主文上訴駁回。上訴審訴訟費用由上訴人負擔。理由請點選下方附件連結以取得判決全文。中華民國106年5月25日最高行政法院第一庭審判長法官劉鑫楨法官胡方新法官程怡怡法官張國勳法官汪漢卿以上正本證明與原本無異中華民國106年5月25日由上訴人負擔。理由請點選下方附件連結以取得判決全文。中華民國106年8月10日最高行政法院第一庭審判長法官劉鑫楨法官胡方新法官程怡怡法官張國勳法官汪漢卿以上正本證明與原本無異中華民國106年8月10日由上訴人負擔。理由請點選下方附件連結以取得判決全文。

III、移除公文模板

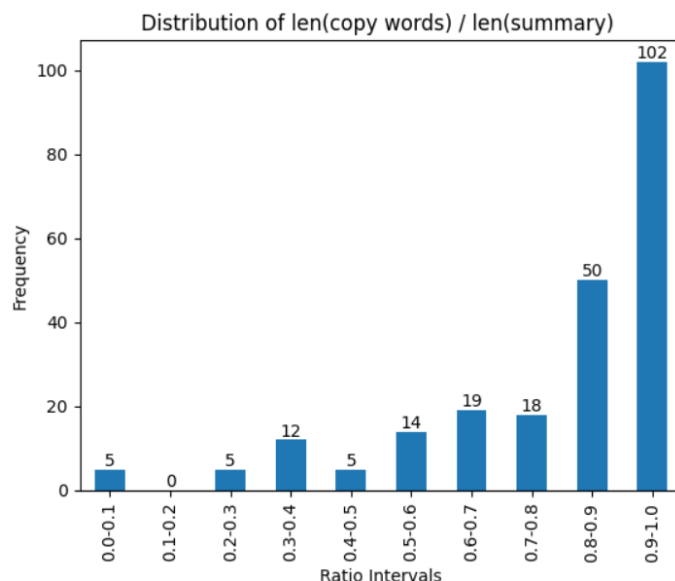
- i. 刪除裁判原文中開頭和結尾前後重複且不必要的部分。

我們發現裁判原文為整份法院判決書中的所有文字，而下圖就是其中一份判決書中開頭、結尾的形式範例，可以看出其有固定格式，並包含了一些重複且不重要的文字。因此我們將裁判原文中「判決如下：」前的文字和「判決如主文：」後的文字刪除，產生一份新的訓練資料。

最 高 行 政 法 院 判 決		政 訴 訟 法 第 255 條 第 1 項、第 98 條 第 1 項 前 段， 判決如主文。	
106 年度判字第 182 號		中 華 民 國 106 年 4 月 13 日	
上 訴 人 日商電化股份有限公司		最高行政法院第一庭	
代 表 人 山本學		審判長法官 劉 鑫 楨	
訴訟代理人 葉信金 專利師		法官 胡 方 新	
被 上 訴 人 經濟部智慧財產局		法官 程 怡 怡	
代 表 人 洪淑敏		法官 張 國 勳	
上列當事人間發明專利申請事件，上訴人對於中華民國 105 年 1 月 21 日 智慧財產法院 104 年度行專訴字第 93 號行政判決，提起上訴，本院 判決 如下：		以 上 正 本 證 明 與 原 本 無 異	
主 文		中 華 民 國 106 年 4 月 13 日	
		書記官 莊 子 誼	

IV、 擷取裁判原文中出現摘要的重點部分

我們分析出有 66% 的摘要，**80% 內容皆從裁判原文中直接整段複製**。如下方的圖表所示，x 軸為摘要文字有出現在裁判原文中的比例，y 軸為資料的筆數，可以發現有 152 筆(66%)的資料 80% 以上的內容是從裁判原文中直接整段複製。



我們利用這個特性，去分析摘要出現在原文中的區段，發現長篇判決的摘要內容多落在後半部，並在經過測試後，得知在判決原文為大於 4000 字的狀況時，摘要內容有高比例出現最後 4000 字，因此我們就**直接擷取其最後 4000 字做為輸入**，大大減少了模型的輸入大小。

2. Model & Training method

I、訓練細節

實驗使用已微調於摘要任務的簡體中文 BART 模型，進一步微調於此法律摘要任務，並根據以下超參數設定訓練過程：

Model Name	IDEA-CCNL/Randeng-BART-139M-SUMMARY
Learning Rate	5e-5
Batch Size	1
Accumulated Step	16
Epochs	30

每筆訓練資料會在原文前方加上”**summary:**”，讓模型與原始訓練方法一致達到更好的效果。

II、生成摘要細節 (Split Chunk)

雖然輸入字數 4000 字相較原本的原文已經縮減很多了，但對於小模型來說，輸入字數仍是過多。我們透過統計發現後段內容通常與摘要高度相關，因此在生成摘要內容時，我們並非採用 sliding window 的想法，而是將摘要分為前 500 字和後 1500 字共四個 500 字的 chunk，分別進行摘要後拼接起來作為最後的摘要。當切 chunk 時可能會有不符合 500 字的規則的情況，我們以至少 250 字作為是否切 chunk 的依據，避免輸入模型的 chunk 內容過少而影響摘要品質。

每個 chunk 再生成摘要過程也需要在前方都加上”**summary:**”，讓生成與訓練的設定一致，我們設定最大摘要長度為 512 最小摘要長度為 200 來限制模型的輸出結果。

3. Analysis

I、錯誤分析與改進

i. Baseline 助教的

按照原助教的程式碼執行，我們發現模型的輸出都非常簡短

而且格式詭異而且非常簡短，並發現最主要的原因為簡體中文的 tokenizer 無法讀取繁體中文的內容，改良原設計成為 ii. 的方法

ii. Baseline with T2S

我們將訓練資料的原文與摘要都轉為簡體中文後訓練摘要模型，在測試階段我們再將生成的簡體中文摘要轉為繁體中文，透過這種方法可以有效地提升 Rouge-2 分數，而且生成的內容也不會詭異且簡短。

iii. Baseline with T2S + data preprocessing

後續經過資料分析我們發現資料中有一些極短的 noise data，而且有統一的公文格式。我們刪除了這些 noise 資料並將公文的前綴與後綴移除，可以幫助我們讀取更多的原文。此外，我們也發現長文的摘要內容常落於摘要的後半段，我們將超過 4000 字的原文僅保留後 4000 字，提供模型更有意義的原文。

iv. Baseline with T2S + data preprocessing + split chunk

再經過更細緻的分析後，我們發現許多摘要內容會將部分原文內容直接抄至摘要，並且出現在原文中後 4000 字的 500 字和後 1500 字。我們設計了第 2 點所描述的 split chunk 的方式，分別進行摘要後連接起來，達到所有方法中的最佳分數。

方法 \ 分數	Rouge-L-P	Rouge-L-R	Rouge-L-F	Rouge-2-P	Rouge-2-R	Rouge-2-F
Baseline	0.7317	0.7838	0.7838	0.1453	0.1521	0.1521
Baseline (with T2S)	0.6695	0.7377	0.7377	0.1678	0.1732	0.1732
Baseline with T2S + data preprocessing	0.6796	0.7248	0.7248	0.1912	0.1937	0.1937

Baseline with T2S + data preprocessing + split chunk	0.6799	0.7428	0.7428	0.2103	0.2172	0.2172
---	--------	--------	--------	---------------	---------------	---------------

II、比較結果

i. 生成結果

藥事法所稱偽藥，係指藥品經稽查或檢驗，有未經覈准擅自製造或所含有效成分之名稱與覈准者不符之情形者，藥事法第 20 條第 1 項第 2 款定有明文。而經中央衛生主管機關明令公告禁止製造、調劑、輸入、販賣或陳列之毒害藥品，若有未經中央衛生主管機關覈准擅自輸入之情形，則屬藥事法人所引起訴法條所稱之禁藥。同法第 22 條第 1、2 款亦定有不明。毒品危害防制條例及懲治走私條例之犯行明確，因而撤銷第一審諭知無罪之判決，經比較新舊法律改判依想像競合犯關係之私運管制物品、輸入私煙，及說明黃騰不另為無罪諭知各部分，均應並予發回。倘上訴理由書狀並未依據卷內訴訟資料，具體指摘，不適用何種法則或如何適用不當，或所指摘原判決違法情事，顯與法律規定得為第三審上訴理由之違法情形，不相適合時，仍應認其上訴為違背法律上之程式予以駁回。二、本件原判決於 105 年 7 月 1 日施行；另藥事法第 88 條亦於 106 年 6 月 14 日修正公佈，並自同年月 16 日施行，原審未及適用新法，此部分案經發回，應注意及之，並此敘明。毒品犯罪之證據，系綜合調查所得之各直接、間接證據而為合理推論。倘其採證認事並未違背經驗法則或論理法則，復已敘述其憑以判斷之心證理由，即不能任意指為違法；又證據之證明力如何，此項自由判斷之職權行使，苟系基於吾人日常生活之經驗，而未違背客觀上應認為確實之定則，又已敘述並敘述其何以為此判斷之理由者，亦不容漫指系違法。而據為適法之第三審上訴理由。一、刑事訴訟法第 163 條已說明調查證據系由當事人主導為原則，法院於當事人主導之證據調查完畢後，認為事實未臻明瞭時，始得斟酌具體個案之情形，予以裁量是否補充介入調查之情形。本件原審審判長於調查證據時，提示法務部調查局對陳家有實施測謊之鑑定書並告以要旨後，經詢問徐○椿及其辯護人有無意見，徒以自己說詞，任意指為違法，或單純為事實上枝節性之爭辯，與法律規定得為第三審上訴理由之違法情形不相適合。

ii. 原始摘要

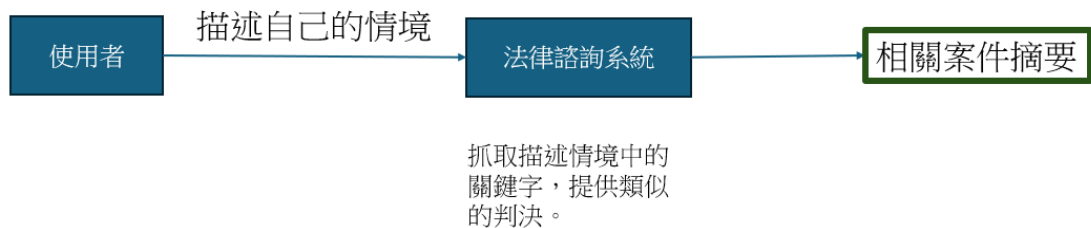
藥事法所稱偽藥，係指藥品經稽查或檢驗，有未經覈准擅自製造，或所含有效成分之名稱與覈准者不符之情形者，藥事法第 20 條第 1、2 款定有明文。而經中央衛生主管機關明令公告禁止製造、調劑、輸入、輸出、販賣或陳列之毒害藥品，或除系旅客或隨交通工具服務人員攜帶自用藥品外，若有未經中央衛生主管機關覈准擅自輸入之情形者，則屬藥事法所稱之禁藥，同法第 22 條第 1 項第 1、2 款亦定有明文。在大陸地區產製藥品，事實上不能經由我國衛生主管機關覈准，更無從予以監督管理，如有將大陸地區產製藥品輸入或攜帶進入臺灣地區，依臺灣地區與大陸地區人民關係條例第 40 條規定既以進口論，其未經覈准擅自輸入者，自應認屬藥事法第 22 條第 1 項第 2 款所稱之禁藥，此為本院一致之見解。原判決依憑證人張○倫證稱黃○騰於報關前，有交代貨櫃內保健食品可能會犯到藥事法等語，且黃○騰亦自承與張○倫間有提到可能有藥事法問題等情，佐以黃○騰與徐○椿於原判決附表(下稱「附表」)3 編號 1 所示物品遭查獲前之通訊內容，並敘明麻黃鹼系行政院公告第四級管制藥品原料藥，兼具偽藥性質，資為認定黃○騰於運送時至少知悉扣案以保健食品名義進口之麻黃鹼系兼具偽藥不能進口之違禁物。然查，原判決理由欄貳、二、乙之(十)既認黃○騰並無運輸第四級毒品先驅原料麻黃鹼之犯意，即認定黃○騰不知徐○椿以保健品名義所託運之物品實為麻黃鹼，卻又以麻黃鹼系兼具偽藥不能進口之違禁物乙節，而認黃○騰有明知偽藥

而輸入之故意,其所載理由前後互相格,難謂無判決理由矛盾之違法。復依首揭說明,藥事法之偽藥與禁藥,核屬相異概念,原判決事實欄一載認黃○騰明知徐○椿以保健品所託運從大陸

4. Proposal（基於 RAG 的法律案件諮詢系統）

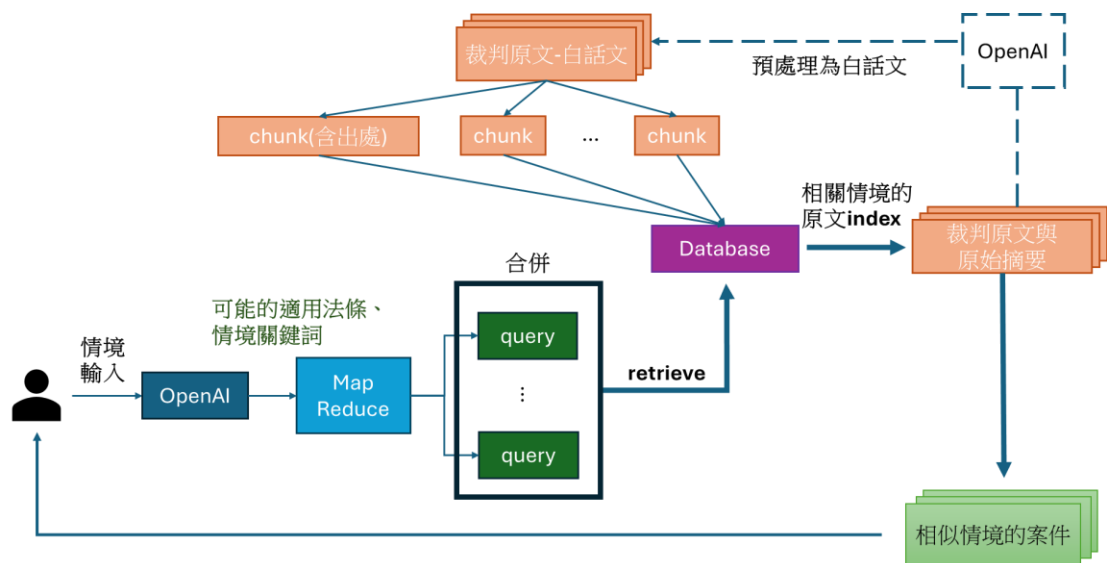
I、簡介

- i. 讓非法律領域的人，能快速找到相關案例。
- ii. 將裁判原文擷取上訴結果：上訴駁回、判決部分廢棄發回高等法院再審、主文原判決廢棄發回高等法院再審……等類型，讓民眾能在上訴前，先確認有無類似這些判決情境後再做上訴，才不會浪費司法資源。



II、法律諮詢系統架構

使用者描述他的情境，GPT-4o 會根據輸入產生更詳細的說明，例如法條與使用情境等等。再放入 MapReduce 產生精準的重點，並去除一些不重要但會影響 RAG 運作的字眼，例如：法、合同、私人……等。將這些重點合併後對已將裁判原文轉換成白話文的 Database 進行檢索（Retrieve），找出前幾個最相似的區塊，並根據這些區塊的來源，回傳相似使用者情境的案件。



III、實作結果

以下是使用者輸入的情境。

```
user_text = "我砍伐了自家前面的森林取用木材，但卻被政府提告。"
```

經由檢索 Database 得出最相關的五個 chunks 後，整理這些 chunks 的來源得到三個 source_id，['111','169','25']，再根據這些 id 從 Database 拿出相似案件的原始摘要與裁判原文。

```
return_text: 1. 在自家前面的森林砍伐木材，即使是自家前面的森林，也需要遵守相關的環境保護規定。若未經相關部門批准，可能會被視為非法砍伐。即使是自家前面的森林，砍伐木材的行為
{'ids': [['doc25_2', 'doc111_1', 'doc111_2', 'doc25_1', 'doc169_1']], 'distances': [[0.36383652687072754, 0.111, '169', '25']]
```

```
=====摘要=====
['sum_doc111']
[{'source': '111'}]
['次按當事人主張有利於己之事實者，就其事實有舉證責任，民事訴訟法第277條前段定有明文，此即舉證責任之分配原則。']
['sum_doc169']
[{'source': '169'}]
['……而所增訂之現行森林法第15條第4項條文，顯係為保障原住民族基本權利，促進原住民族生存發展，並尊重原住民族之']
['sum_doc25']
[{'source': '25'}]
['查犯104年5月6日修正公布施行（同月8日生效）前森林法第52條第1項竊取森林主產物罪者，除處六月以上五年以下有期徒刑']
```

```
=====裁判原文=====
['ori_doc111']
[{'source': '111'}]
['最高法院民事判決一〇六年度台上字第三三號上訴人游木錡（原名游江河）林錫坤共同訴訟代理人羅豐胤律師蘇仙宜律師初']
['ori_doc169']
[{'source': '169'}]
['最高法院刑事判決106年度台上字第37號上訴人伍彥章選任辯護人呂秀梅律師上訴人伍添富全致賢上列上訴人等因違反森林']
['ori_doc25']
[{'source': '25'}]
['最高法院刑事判決一〇六年度台上字第七七三號上訴人江建和選任辯護人郭美春律師蔡瑜軒律師上訴人陳宏鑫（原名陳佳祚']
```

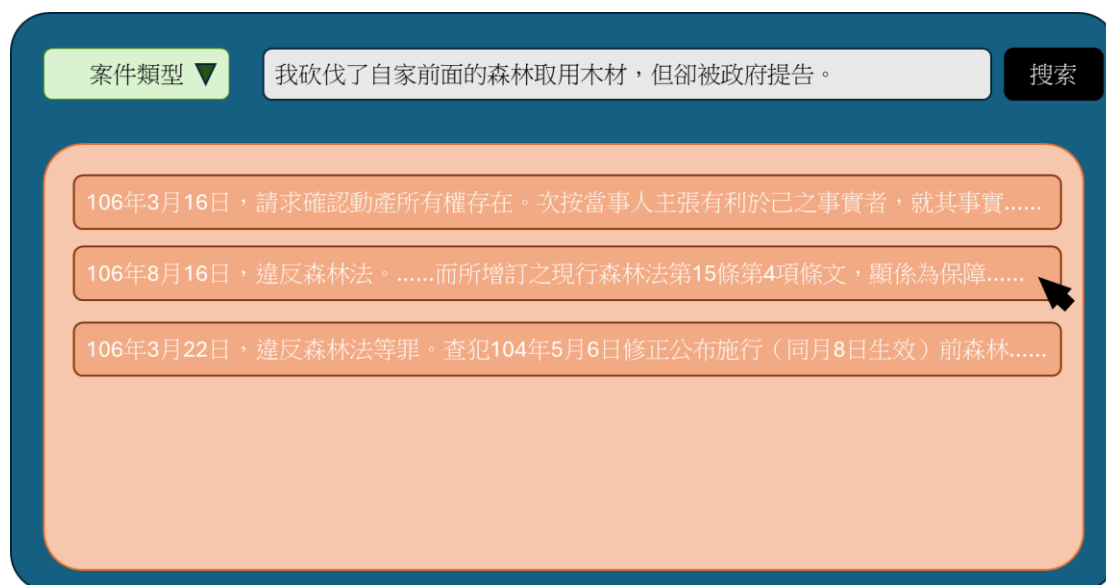

IV、預期網頁的示意圖

i. 使用者輸入及選擇案件類型（可不選）



The interface shows a search bar with the placeholder text "請描述你的情境..." and a "搜索" (Search) button. To the left, there is a dropdown menu labeled "案件類型" (Case Type) with a downward arrow. Below the dropdown, three buttons are visible: "# 刑事法" (Criminal Law), "# 民事法" (Civil Law), and "# 公法" (Public Law). The entire interface is enclosed in a blue border.

ii. 得出最相關的案件



The interface shows the search bar with the text "我砍伐了自家前面的森林取用木材，但卻被政府提告。" (I cut down the forest in front of my house to use wood, but I was sued by the government.) and the "搜索" (Search) button. Below the search bar, there is a list of three search results, each in a light orange box:

- 106年3月16日，請求確認動產所有權存在。次按當事人主張有利於己之事實者，就其事實.....
- 106年8月16日，違反森林法。.....而所增訂之現行森林法第15條第4項條文，顯係為保障.....
- 106年3月22日，違反森林法等罪。查犯104年5月6日修正公布施行（同月8日生效）前森林.....

The interface is enclosed in a blue border.

iii. 顯示該案件的摘要與原文

106年8月16日

違反森林法

#刑事法

摘要

.....而所增訂之現行森林法第15條第4項條文，顯係為保障原住民族基本權利，促進原住民族生存發展，並尊重原住民族之傳統習俗、文化及價值觀而設，且明文規定：係與「採取森林產物」有關之事項，與行政院原函請立法院審議之上述森林法第15條第4項原提案條文之草案立法說明所稱之採取雜草、枯枝、落葉及許可證等，完全無涉。又依前揭森林法第15條第3項規定授權訂定之「國有林林產物處分規則」，係著重於國有林林產物處分之管理，以落實林業永續經營之政策，乃屬於採取國有林林產物之一般規定，該規則雖於第14條第1項第7款、第8款就原住民之採取國有林林產物，設有專案核准之規定，另對「採取副產物或藥用林產物者」、「打撈漂流竹木者」，亦於同條項第10款、第12款設有專案核准之規定

106年8月16日

違反森林法

#刑事法

原文

最高法院刑事判決106年度台上字第37號上訴人伍彥章選任辯護人呂秀梅律師上訴人伍添富全致賢上列上訴人等因違反森林法案件，不服臺灣高等法院臺中分院中華民國104年10月8日第二審判決（104年度原上訴字第25號，起訴案號：臺灣南投地方法院檢察署102年度偵字第4205號），提起上訴，本院判決如下：主文原判決關於違反森林法部分撤銷，發回臺灣高等法院臺中分院。理由一、本件原判決認定上訴人伍彥章、伍添富、全致賢有原判決事實欄所載之共同結夥2人以上竊取森林副產物桑黃，並為搬運贓物使用車輛之犯行，甚為明確，因而維持第一審適用民國104年5月6日修正公布施行前之森林法第52條第1項第4款、第6款規定，論處其3人結夥2人以上竊取森林副產物，