伍、分析與結論 V. Analysis and Conclusion  
 由於時間以及計算資源不足，我們將大量的改進都放在第一部分(Document Retrieval)，以下是我們Document Retrieval部分更改的方法及結果：一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

首先CKIP（Chinese Knowledge and Information Processing）是一個針對中文語言處理的工具包和系統。它由台灣中研院資訊科學研究所語言學研究群所開發，我們運用CKIP進行斷詞，再由hanlp進行詞性標注，我們也有調整wiki page return的數量，嘗試了很多次後發現從初始值 5 調高到約15以後recall上升幅度漸漸趨於0，接下來我們也有嘗試HW3裡面使用的Jieba來斷詞，其中我們發現Jieba還有許多其他功能或模式，像是搜尋引擎模式：找出所有可能的組合，以便於搜尋引擎建立關鍵字，也適用於文本分析，我們也嘗試了Jieba的詞性標註工具，Jieba和hanlp的差別在於模型訓練方法、詞性標註的準確性、分詞等等，由結果可以推測hanlp的效果在這個任務上是優於jieba的。最後是使用Hw3使用的TF-IDF方法，其結果是所有嘗試方法中最好的。

第二部分結果如下：

Training scores :

{'F1 score': 0.44885298281633573, 'Precision': 0.3387890625000051, 'Recall': 0.66484375}

Start validation:

Validation scores => {'F1 score': 0.4433403573394844, 'Precision': 0.33393229166666527, 'Recall': 0.659375}

第三部分結果如下：

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

第二部分及第三部分我們沒什麼做更改，使用的是原本的bert-base-chinese，它是一種基於Transformer的NLP模型。bert-base-chinese是基於BERT衍生出的模型，專門針對中文進行訓練和處理。

由於每有算好時間導致好多東西都來不及弄真的滿可惜的，總結一下如果還有時間的話會想嘗試的方式：

1. TF-IDF的部分還可以做得更好。
2. 第二部分好像還有資料不平衡的問題需要處理，以及可以透過分類句子長度來進行簡單分類。
3. 聽了其他組的報告後也發現我們有滿多可以改進的地方，之後可能會想要嘗試多訓練幾個其他的模型，並且把結果ensemble起來。