法規資料搜尋引擎

組員：黃少麒 莊東翰 杜欣洋　指導老師：吳宜鴻 老師

特色

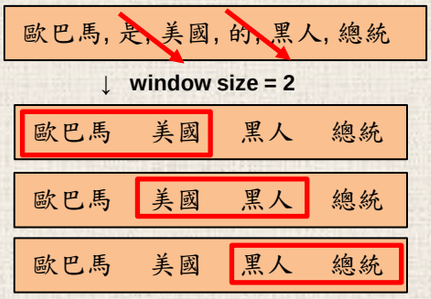
此系統有兩大特色，最大特色就是在現有的法規網站「全國法規資料庫」中沒有的搜尋方式為「口語化搜尋」，使用者可不用透過法律中的專業用語去尋找想要的答案，可用一般對話時會講的句子去搜尋。另一個特色為「以條文形式輸出」，現有的法規網站「全國法規資料庫」為顯示法規名稱，但一般民眾在使用時可能會不知道裡面詳細的法規內容是什麼，需要逐一尋找才能找到想要的結果，相當耗時，所以我們以列表形式呈現結果，讓使用者可以一目了然的找到答案。

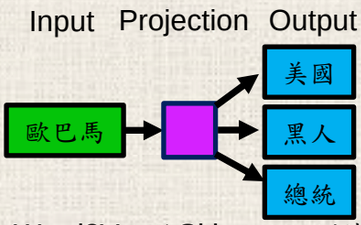
未來展望

1. 將所有現有法規擴充至我們的系統中供使用者查詢。
2. 擴充案例搜尋讓使用者可以查詢到一些法官已經判決的案例，或是其他使用者提供的案例來做比對。
3. 與一些律師事務所合作建立對答機器人讓使用者可以直接在對話窗詢問在線律師。

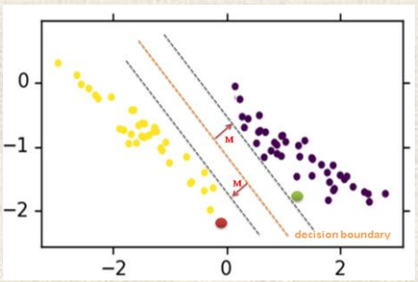
方法

獲得使用者輸入之句子或關鍵字後使用Jieba斷詞與去除停詞的方式去獲取句子之重要詞彙，之後透過Word2Vec的Skip-gram模型尋找中心詞對應出的更多詞彙，此步驟可用口語化詞彙當中心詞彙對應到許多專業用語，以達成口語化之目的。接著這些詞彙透過TF-IDF算出每個小類別的分數(feature)，以許多feature組成的向量餵給SVM (Support Vector Machine支援向量機)以用來找這些向量的邊界。最後SVM得到此問題的小分類後去資料庫中搜尋屬於此小分類的法規。

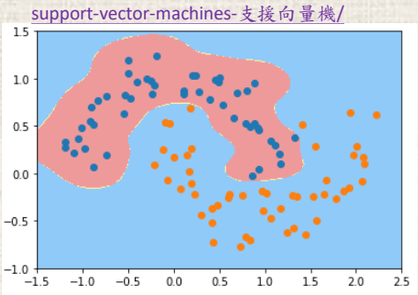




Word2Vec的Skip-gram訓練模型與window size紀錄詞關係之範圍



https://chtseng.wordpress.com/2017/02/04/support-vector-machines-支援向量機/



https://www.itread01.com/content/1534167680.html

SVM模型尋找兩組dataset的邊界（線性與非線性圖形）

結果

左圖為全國法規資料庫中輸入較為口語化之句子的搜尋結果，右圖為我們的法規資料搜尋引擎輸入較口語化之句子的結果，可以很明顯的發現全國法規資料庫對於口語化句子是沒有辦法處理的。