Course IR

homework 1

P76071200 馬崇堯

一.使用工具、技術：

1.程式語言 - Python
2.深度學習框架 - Keras
2-1 詞轉詞性工具 - Spacy
3.GUI呈現 - Web後端Django
- Web前端Html
4.cElementTree - parser xml
5.Python bulid-in json parser

二.資料前處理：

1.讀檔並對副檔名做判別
2.根據不同檔案類型做parser
3-1.xml對內文和抬頭做分詞和分句
3-2.json對內文做分詞和分句
4.建立字、句子、文章的關聯表
5.關聯表建進mysql database

三.資料統計：

1.字元、字詞都在分詞時一併計算
2.句數先以句點分句計算

四.檢索：

1.前端將要搜尋的字傳至後端
2.後端接收後去資料庫查表
3.將資料庫中資料撈出，處理後顯示自前端

五.End of sentence：

由pubmed上爬文章並以句點為標準來分句，再進行label後將每個分句經由 Spacy轉成相對應得詞性。再將詞性轉成向量後餵進keras中的LSTM模型再經由全連接層後輸出成0或1。藉此來判斷是否為一個完整的句子結構。

六.問題討論及需改進部分：

1.在前端呈現的方式不夠一目了然。
2.目前僅有讀檔功能，需將資料事先放進相對應的資料夾。但理想上應該是可以直接坐到檔案上傳至後端的功能。
3.有時候還是會檢索到完全不相干的句子，尤其是在檢索很頻繁出現的詞，要重新審視演算法。
4.預測EOS的模型，表現差強人意，要考慮是否訓練資料不足或是訓練方法上有問題。