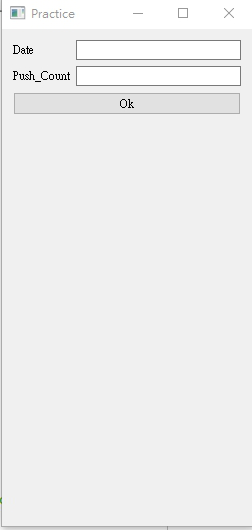
**1.爬文章.py**

**說明：輸入日期即想要多少推問數以上的文章，將其以以html的形式個別存成一個個檔案整個爬下來**

須改動的部份：

42行 將想要爬的ptt的版的網址輸入

201行 將要貯存文章的位置輸入(c\_folder改動位置在”相同位置.ini”)，文章便會一個一個存下來



**2.整理成文章.py**

**說明：(將多餘html的去除只留下想要的內容)**

**60~61 將文章內容留下**

須改動的部份：

12行 輸入第一個程式碼爬下來的txt檔所在資料夾位置(c\_folder改動位置在”相同位置.ini”)

41~54行將所要發文者相關資訊留下(例如：ID、版類、標題、發文時間)

67~87行將評論者相關資訊留下(例如：ID、推噓、留言時間、留言內容)

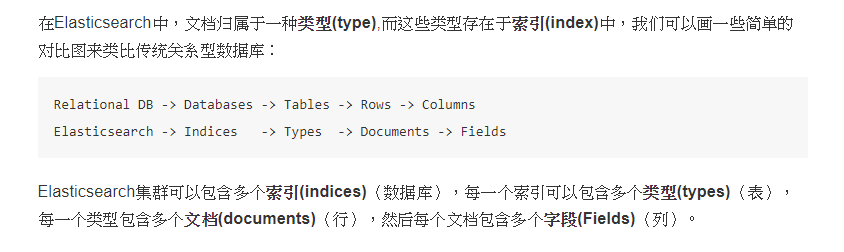
96行選擇寫入的檔案(將同一天的所有文章匯入同個txt檔以便匯入es)

**3-1.匯入ela.py**

須改動的部份：

10行輸入文章所在資料夾位置(h\_folder改動位置在”相同位置.ini”)

21行 輸入index type 便可將文章匯入es中



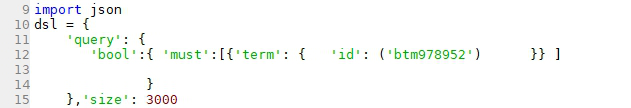
**3-2.搜尋ela裡的文章.py**

**說明：9行~15行輸入欲查詢之內容**

須改動的部份：

12：欲搜尋之關鍵字

15：所搜尋之最大數量



17行 body 為欲貯存之搜尋內容(例如：只想要文章內容)



19行 輸入欲寫入之檔案名(一個檔案)( content\_file改動位置在”相同位置.ini”)

**4.分離文章.py**

**說明：經由搜尋後同個作者之文章會放在同個檔案中**

**故將同個作者的文章分離成一個一個檔案**

須改動的部份：

7行輸入文章作者及完成( content\_file改動位置在”相同位置.ini”)

**5.Tokenlize**

**程式說明：**

**11-69以jieba.add\_word()加詞進結巴詞庫**

**85-103用結巴切文章並存到指定的資料夾**

須改動的部份：

106-113輸入要切的文檔所在的資料夾以及切好的文檔要放到哪個資料夾(訓練集與測試集已手動分好)( beforetest\_folder、beforetrain\_folder、aftertest\_folder、aftertrain\_folder改動位置在”相同位置.ini”)

**6.Evaluate**

**程式主要功能：建model及預測**

**程式說明：**

**73-78選擇要用的model(參數也在此調整)**

**91-102將欲預測的文章用結巴切好並預測後輸出結果**

須改動的部份：

51輸入訓練集所在資料夾(aftertrain\_folder改動位置在”相同位置.ini”)

89輸入測試集所在資料夾(aftertest\_folder改動位置在”相同位置.ini”)