**10/11-10/4**

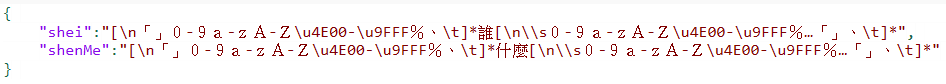
\*\*\*\*\* **語料前處理**

1. 收集中研院平衡語料庫中包含「誰」及「什麼」的語料

首先，將中研院平衡語料庫涵蓋了「誰」和「什麼」的語料完整地儲存至文本文件中，並利用一套程式模型將初步收集到的語料進行預處理的動作。

使用中研院現代漢語平衡語料庫 (https://asbc.iis.sinica.edu.tw/) 中與中文wh疑問詞相關的語料，其中包括2979筆含有「誰」的語料和10671筆含有「什麼」的語料。經語料預處理過後抽取出含有「誰」及「什麼」的語料分別為2695 及 10391筆。

1. 撰寫Regex正規表示式以擷取包含「誰」和「什麼」的語句：

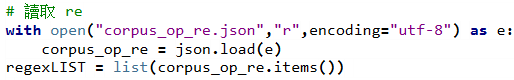


(Regex 含意)

抽取 「\n (換行)、「」、0-9 (數字)、a-z (小寫字母)、A-Z (大寫字母)、\u4E00-\u9FFF% (所有漢字)、\t (tab)」中的「誰」

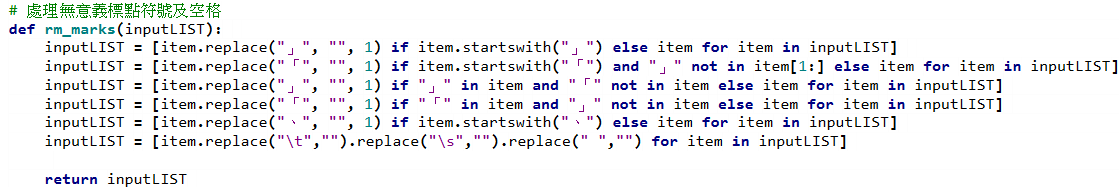
此 .json檔包含Regex正規表示式，旨在匹配包含「誰」和「什麼」的句子。這些模式考慮了各種字元和符號，包括數字、字母、中文字元、標點符號等，以涵蓋文本中可能存在的不同字元和排版方式，以確保匹配的準確性和全面性。目的是在文本中識別並提取包含目標詞彙的部分，以便進行進一步的文字處理和分析，在應用大規模語料庫時，提高文字處理的效率和準確性。

1. 導入Regex正規表示式模組和 .json文件



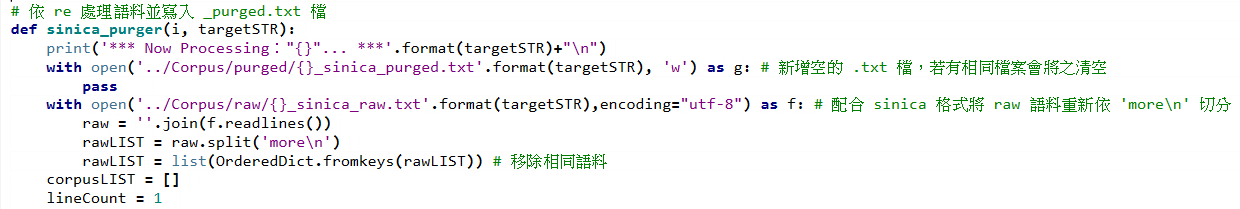
導入了Regex正規表示式模組和.json文件，以進一步處理平衡語料庫的排版問題。

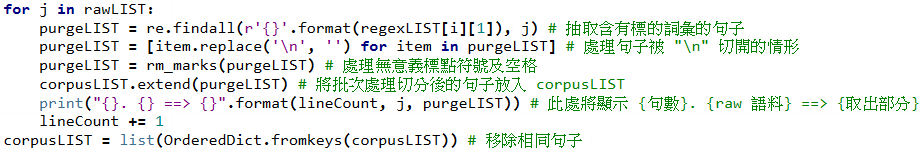
1. 整理語句排版

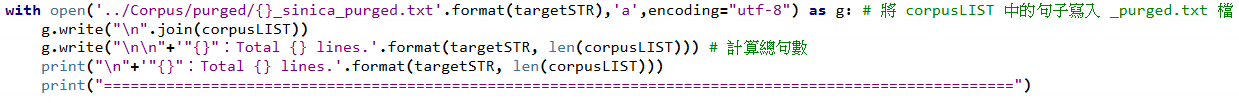


將多餘的標點符號及空格處理掉，讓語料看起來乾淨。

1. 將篩選出的包含「誰」和「什麼」的語句整理並儲存到新的文字檔案中：



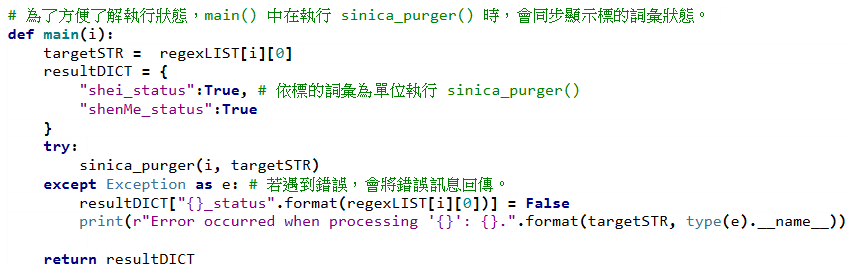




這段Python程式碼處理文字語料，它採用Regex正規表示式來抽取文字的特定部分，並將抽取結果記錄到新的文字檔案中。

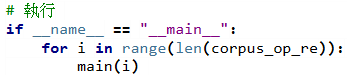
這包括處理句子被分行切開的情況、移除重複的語料、並且配合sinaca格式，將原始語料以”more\n”切分、應用先前處理無意義符號及空格之function (rm\_marks)。並將批次處理切分過後，包含目標詞彙的句子抽取出來放入corpusLIST，計算總句數，並放入新建立的文字檔案中。除此之外，主程式會輸出原始語料的句子數量、原始完整語料以及最終處理過後的語料，以便進行進一步比對和檢查。

1. 檢查執行狀態



在執行sinica\_purger 時，為了確認執行狀態，會同步標示標的詞彙的狀態。檢查過程依標的辭彙為單位，檢查標的辭彙的狀況。若是有錯誤，將錯誤訊息回傳以便進一步檢查。

1. 執行程式



此段程式碼為程式進入點，執行以上程式碼，達到語料預處理的效果。