Introduction-to-Information-Retrieval-and-Text-Mining, HW 1

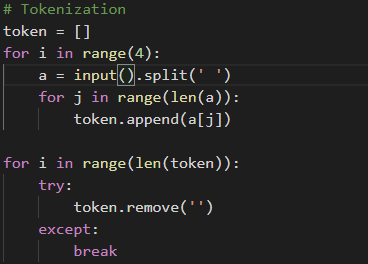
B07302230, 財金三 王博奕

1. 執行環境：Visual Studio Code
2. 程式語言：Python 3.8
3. 執行方式：
   1. Pip install nltk以便於接下來使用Porter’s algorithm。
   2. 輸入本次作業之文本
4. 作業處理邏輯說明：

根據作業說明所提供之指示，本次作業主要分為五個步驟，以下將依序說明。

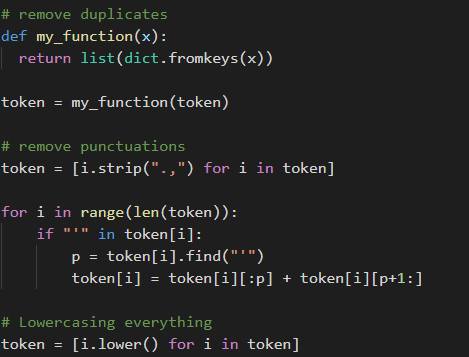
1. Tokenization

首先將資料讀進程式中，接著用空格來分割每個詞成為token。



1. Lowercasing everything

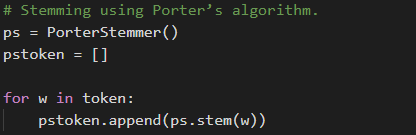
在將token變為小寫前，我先對我的資料做了兩件事情，第一為刪除list中重複的element，第二為刪除部分element存在標點符號的問題。完成上述兩步驟最後再小寫化所有的token。



1. Stemming using Porter’s algorithm

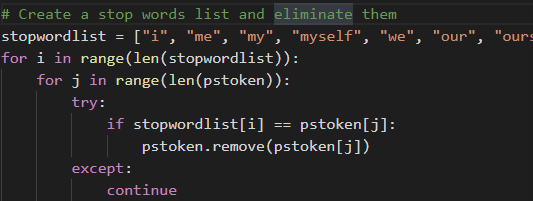
由於先前已import完成，因此即可使用套件來執行Porter’s algorithm。





1. Stopword removal

由於作業要求不能引用套件，因此我直接參考網路上寫好的list在自己的程式碼建一個list (由於list過長因此僅截圖部分)



1. Save the result as a txt file

完成上述的步驟即可輸出為一個txt檔。



以下為最後結果。

