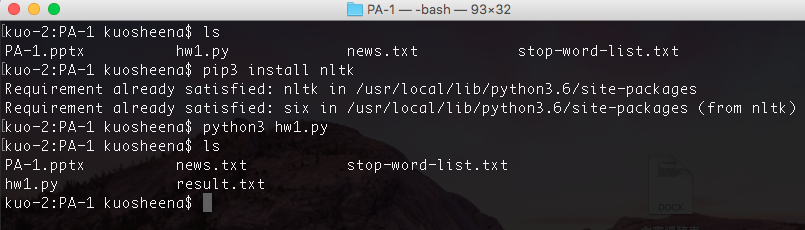
PA-1 R05725060郭欣宜

執行環境：Terminal

程式語言：Python3

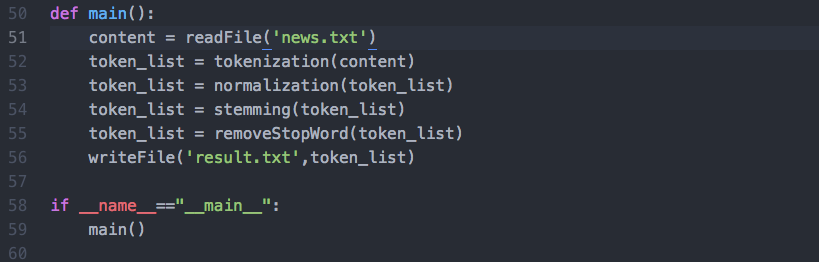
執行方式：

1. pip3 install nltk
2. python3 hw1.py



作業邏輯處理說明：

1. 先讀進news.txt這個file
2. 把file的content 進行tokenization存到token\_list
3. 之後依序對token\_list進行stemming、remove stop word後再存結果到result.txt這個file



readFile:

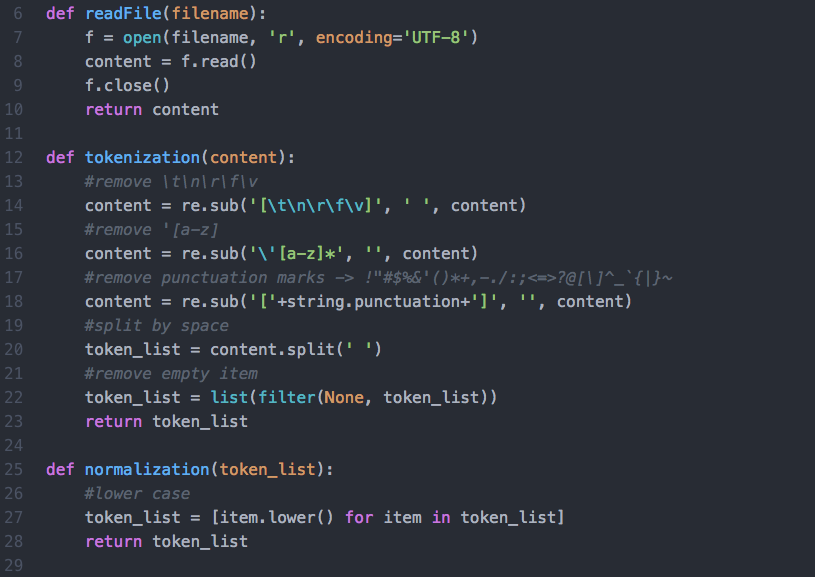
1. 把news.txt的content讀進來

tokenization有進行五項處理：

1. 把content的換行、tab等轉換為space
2. 把’後面為小寫的字刪除(example: that’s -> that)
3. 去除標點符號
4. 利用space切token存進token\_list
5. 把token\_list中空的token去除

normalization:

1. 把所有token都變為小寫的string



stemming:

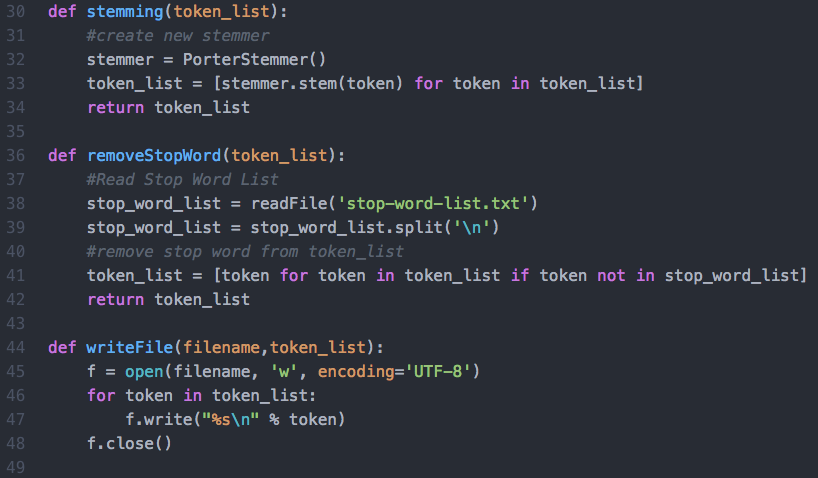
1. 建立nltk的PorterStemmer到stemmer
2. 針對每個token進行stemming

removeStopWord:

1. 讀stop-word-list.txt的file，並利用換行切出stop word
2. 如果token不是stop word 才保存下來

writeFile:

1. 最後把token的結果寫到result.txt



result.txt

