# HW1 Report

R07725021 資管碩1 洪靖雯

## 1.執行環境

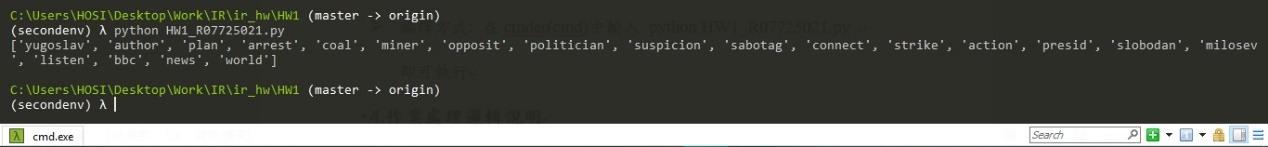
* Sublime
* Cmder

## 2.程式語言

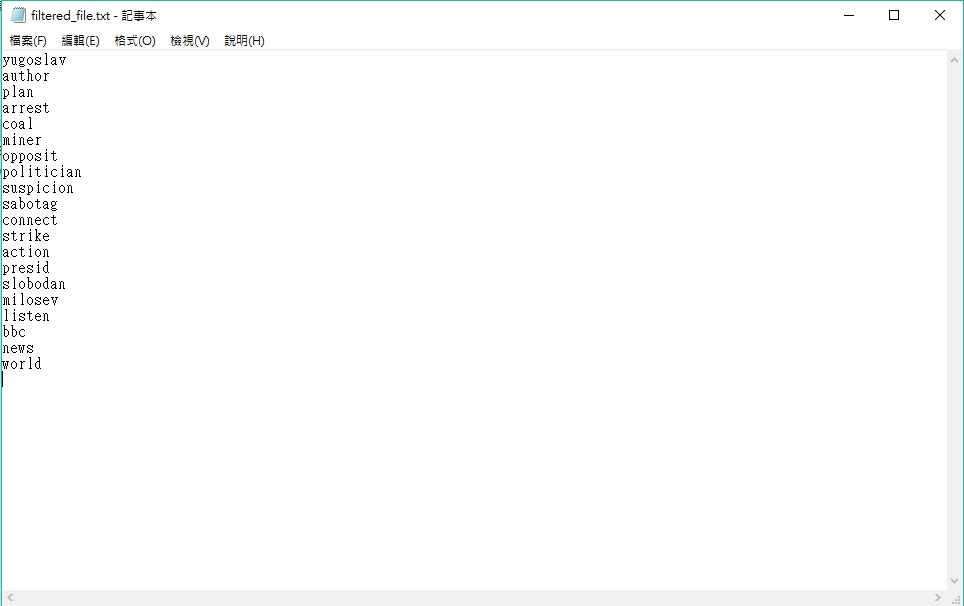
* Python 3.6.6

## 3.執行方式

* 套件: requests、bs4、nltk
  + 有另外用conda 4.5.8架虛擬環境安裝，直接用在system裡應該也可行
  + 使用pip install方式安裝
    - pip install requests
    - pip install bs4
    - pip install nltk
* 編譯方式: 在cmder(cmd)中輸入 python HW1\_R07725021.py

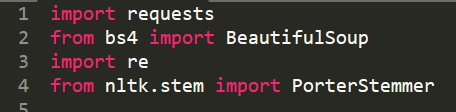
即可執行

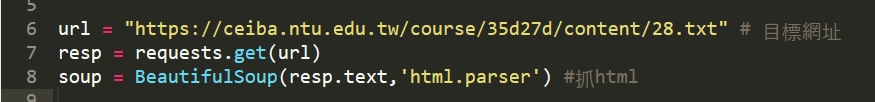
* 輸出結果 (輸出檔名: filtered\_file.txt)

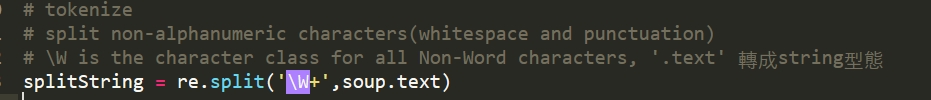


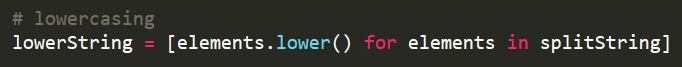
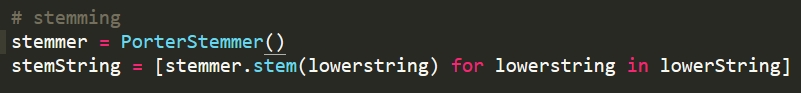
## 4.作業處理邏輯說明

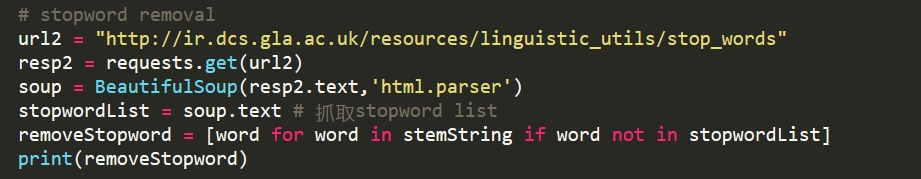
* Import需要的套件



* 先將目標文件([English news Document](https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt))爬蟲下來
* Tokenization
  + 去除non-alphanumeric characters (空白與符號)



* Lowercasing
* Stemming (Porter’s algorithm)
* Stopword removal
* 先將老師提供的stopword list爬蟲下來
* 再跟前面處理好的list做比對，如果該char在stopword list，表示為stopword => 去除



* Save result as txt file

