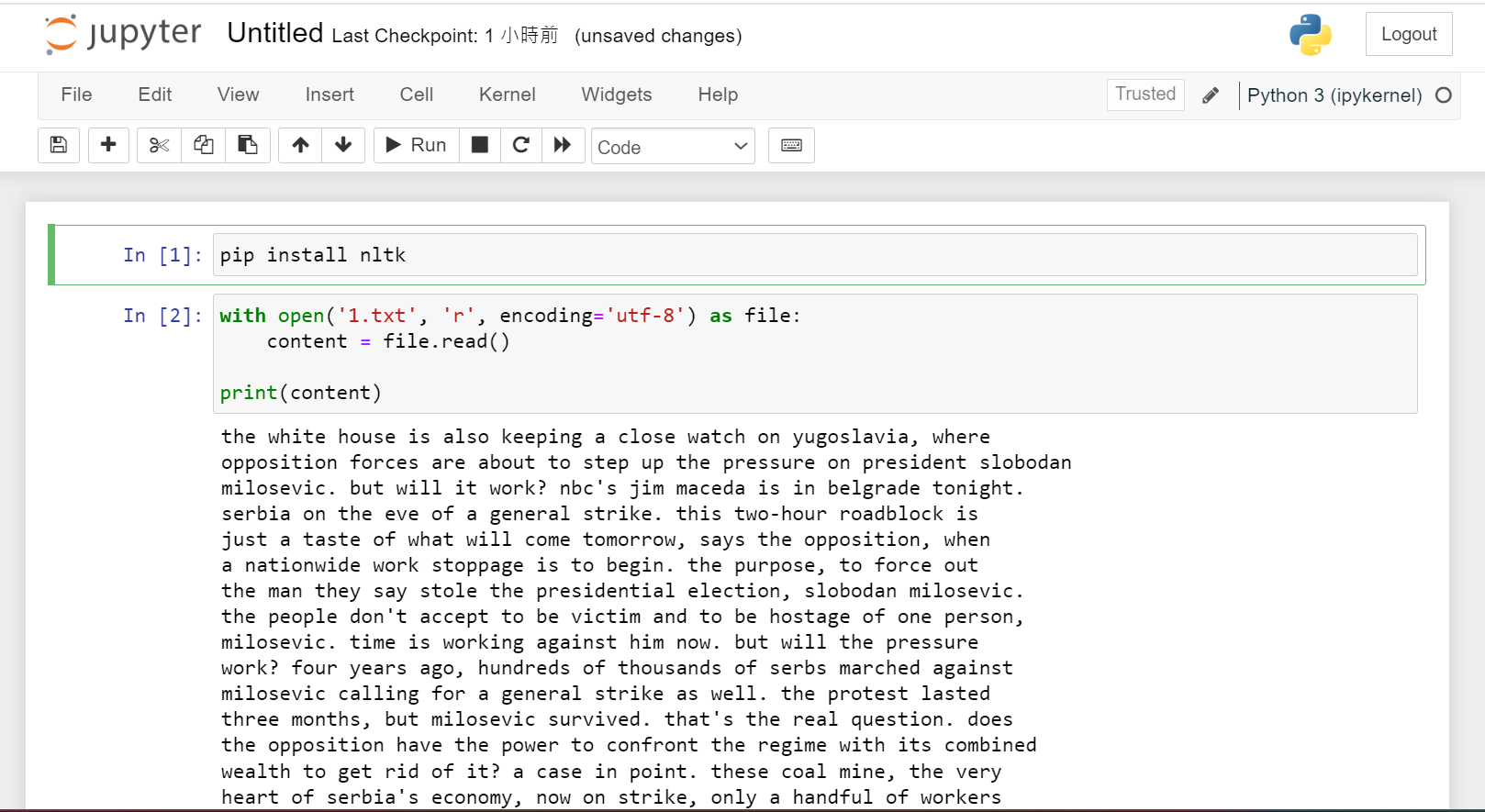
文字探勘HW1

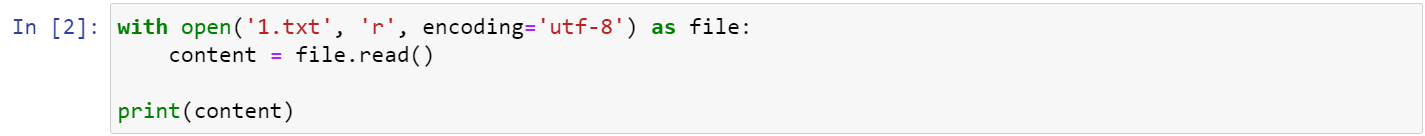
資管三 B11705033 江睿宸

1. 執行環境：Jupyter Notebook
2. 程式語言：Python3
3. 

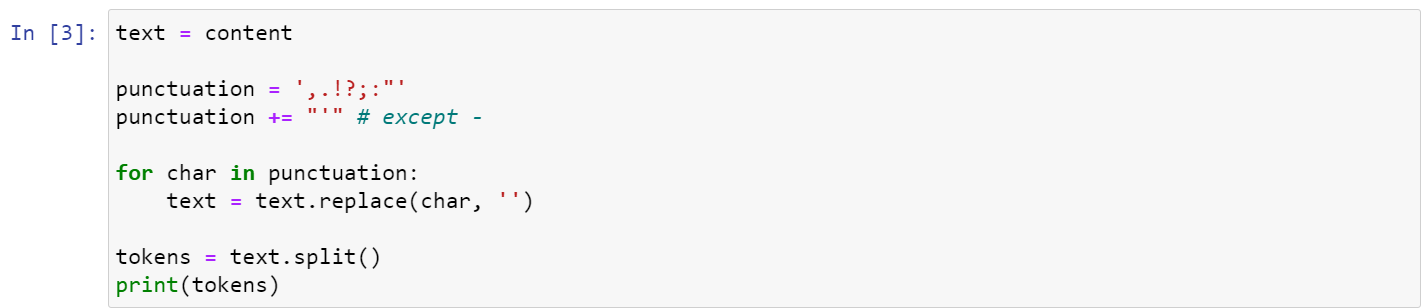
如果沒有ntlk套件需要先安裝(執行第一個cell)

使用jupyter notebook執行全部

1. 讀入檔案



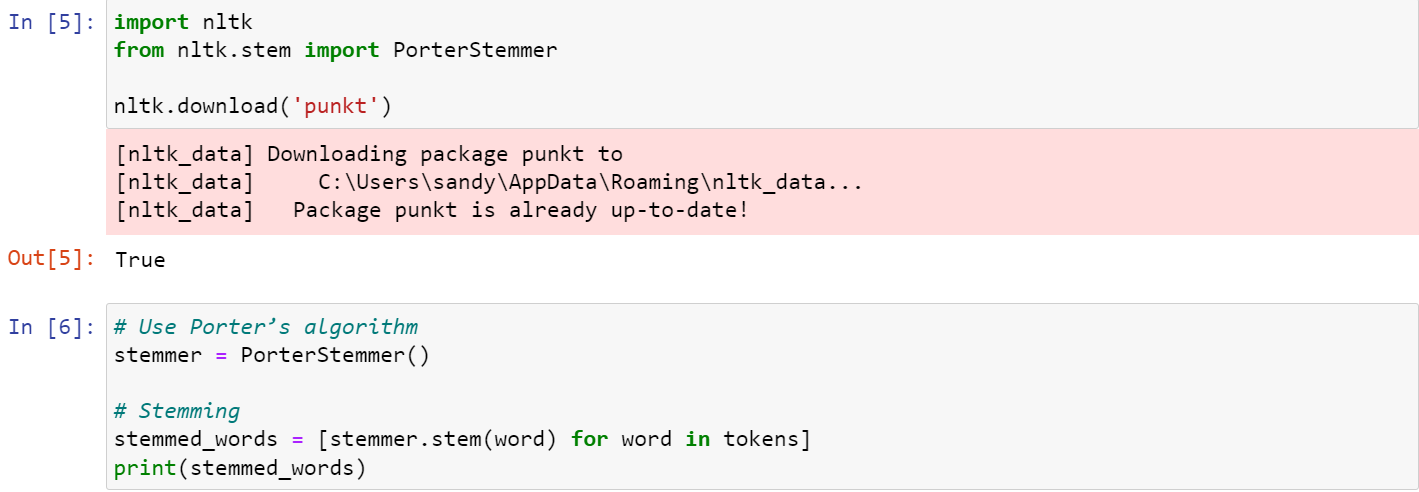
Tokenize：將標點符號去除後，將單字沿空格處分開。Hyphen兩端視為同一單字，故不進行去除。



Lowercasing：使用.lower()進行lowercase



Stemming using Porter’s algorithm：使用ntlk套件進行Porter’s algorithm，完成Stemming



Stopword removal：使用 [**NLTK's list of english stopwords**](https://gist.github.com/sebleier/554280#file-nltk-s-list-of-english-stopwords)將stopwords去除

