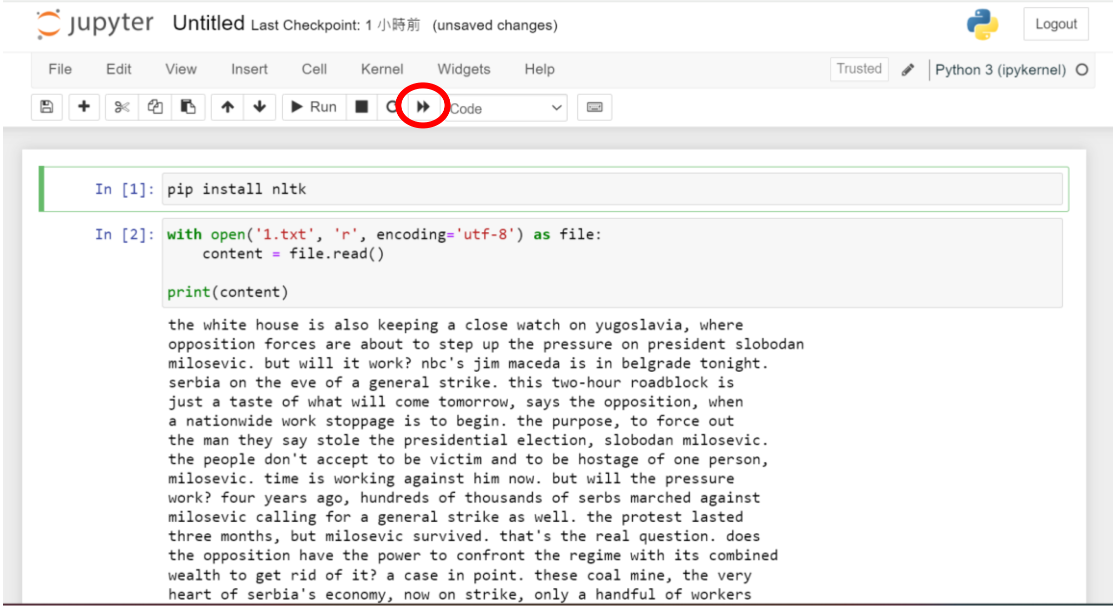
文字探勘 HW1

資管三 B11705033 江睿宸

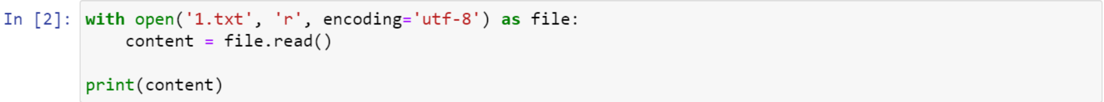
1. 執行環境：Jupyter Notebook
2. 程式語言：Python3



```
In [1]: pip install nltk

In [2]: with open('1.txt', 'r', encoding='utf-8') as file:
            content = file.read()

        print(content)

        the white house is also keeping a close watch on yugoslavia, where
        opposition forces are about to step up the pressure on president slobodan
        milosevic. but will it work? nbc's jim maceda is in belgrade tonight.
        serbia on the eve of a general strike. this two-hour roadblock is
        just a taste of what will come tomorrow, says the opposition, when
        a nationwide work stoppage is to begin. the purpose, to force out
        the man they say stole the presidential election, slobodan milosevic.
        the people don't accept to be victim and to be hostage of one person,
        milosevic. time is working against him now. but will the pressure
        work? four years ago, hundreds of thousands of serbs marched against
        milosevic calling for a general strike as well. the protest lasted
        three months, but milosevic survived. that's the real question. does
        the opposition have the power to confront the regime with its combined
        wealth to get rid of it? a case in point. these coal mine, the very
        heart of serbia's economy, now on strike, only a handful of workers
```

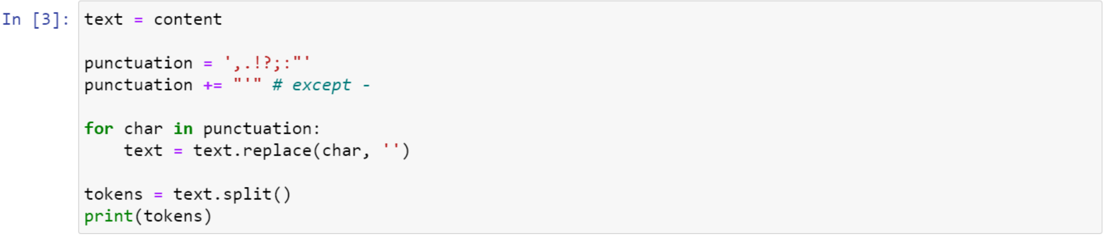
3. 如果沒有 ntlk 套件需要先安裝(執行第一個 cell)

   使用 jupyter notebook 執行全部
4. 讀入檔案



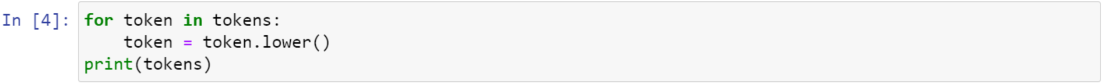
```
In [2]: with open('1.txt', 'r', encoding='utf-8') as file:
            content = file.read()

        print(content)
```

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。



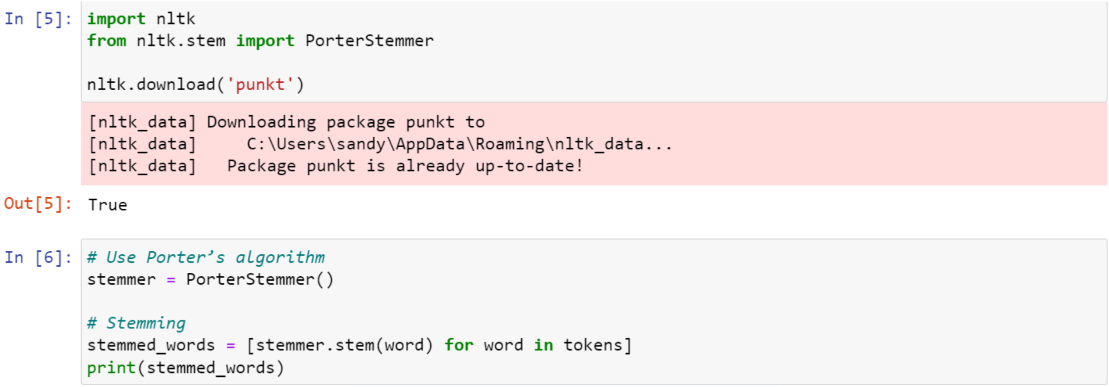
```
In [3]: text = content

        punctuation = ',.!?;:"'
        punctuation += "'" # except -

        for char in punctuation:
            text = text.replace(char, '')

        tokens = text.split()
        print(tokens)
```

Lowercasing：使用.lower()進行 lowercase



```
In [4]: for token in tokens:
            token = token.lower()
        print(tokens)
```

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



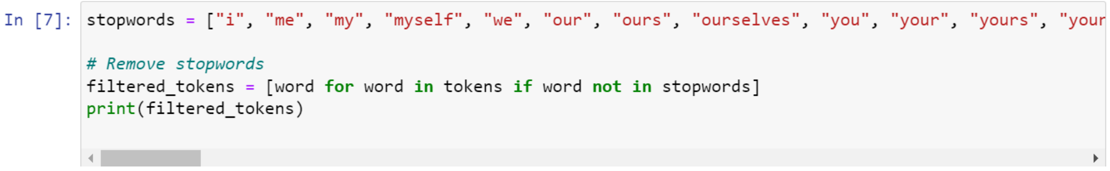
```
In [5]: import nltk
        from nltk.stem import PorterStemmer

        nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\sandy\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
Out[5]: True
```

```
In [6]: # Use Porter’s algorithm
        stemmer = PorterStemmer()

        # Stemming
        stemmed_words = [stemmer.stem(word) for word in tokens]
        print(stemmed_words)
```

Stopword removal：使用 NLTK's list of english stopwords 將 stopwords 去除



```
In [7]: stopwords = ["i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "your

        # Remove stopwords
        filtered_tokens = [word for word in tokens if word not in stopwords]
        print(filtered_tokens)
```