資訊檢索與文字探勘導論
HOMEWORK 1

資管三吳驊祐 B09705009

2022-09-23

1 Environment

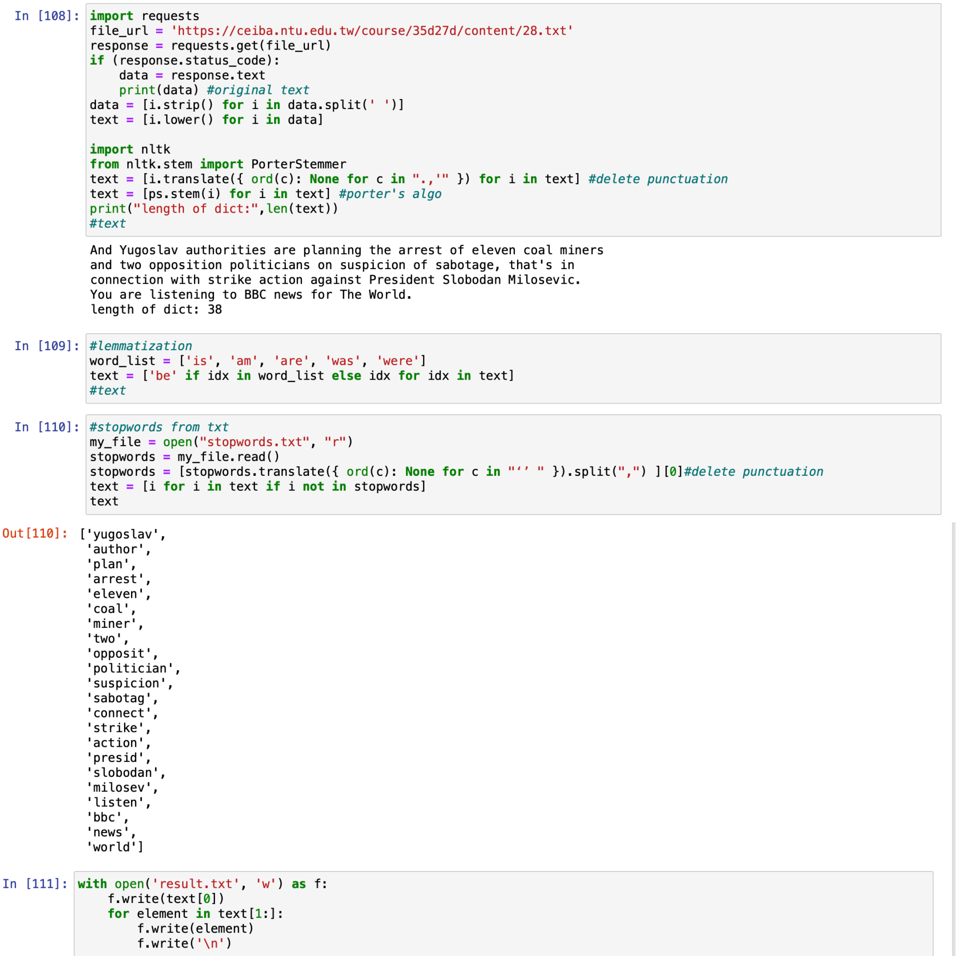
Using Jupyter Noteook

2 Language

Python3

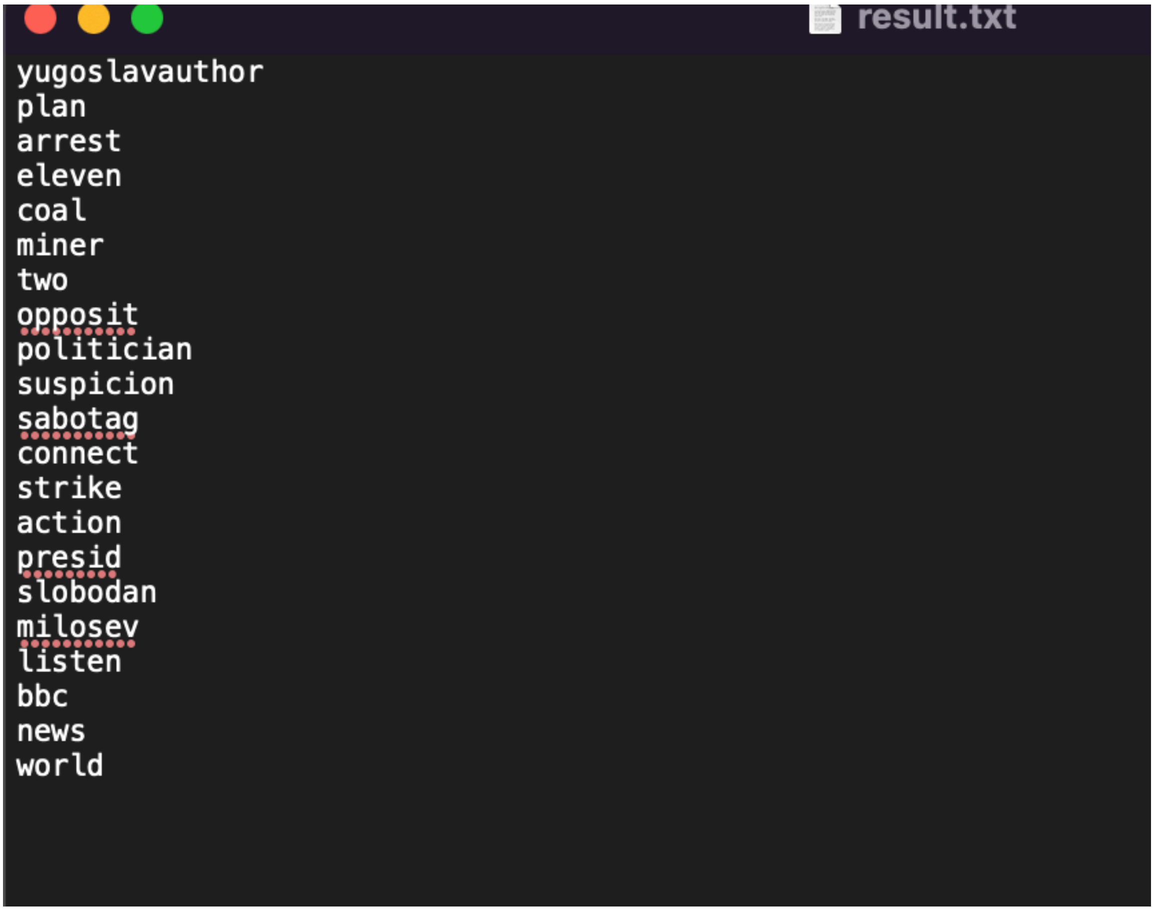
3 Execute

`import requests, from nltk import` *`PorterStemmer`* -> Use jupyter notebook to run the ipynb file.



```
In [108]: import requests
          file_url = 'https://ceiba.ntu.edu.tw/course/35d27d/content/28.txt'
          response = requests.get(file_url)
          if (response.status_code):
              data = response.text
              print(data) #original text
          data = [i.strip() for i in data.split(' ')]
          text = [i.lower() for i in data]

          import nltk
          from nltk.stem import PorterStemmer
          text = [i.translate({ ord(c): None for c in ".,'" }) for i in text] #delete punctuation
          text = [ps.stem(i) for i in text] #porter's algo
          print("length of dict:",len(text))
          #text

          And Yugoslav authorities are planning the arrest of eleven coal miners
          and two opposition politicians on suspicion of sabotage, that's in
          connection with strike action against President Slobodan Milosevic.
          You are listening to BBC news for The World.
          length of dict: 38

In [109]: #lemmatization
          word_list = ['is', 'am', 'are', 'was', 'were']
          text = ['be' if idx in word_list else idx for idx in text]
          #text

In [110]: #stopwords from txt
          my_file = open("stopwords.txt", "r")
          stopwords = my_file.read()
          stopwords = [stopwords.translate({ ord(c): None for c in "‘’ " }).split(",") ][0]#delete punctuation
          text = [i for i in text if i not in stopwords]
          text

Out[110]: ['yugoslav',
           'author',
           'plan',
           'arrest',
           'eleven',
           'coal',
           'miner',
           'two',
           'opposit',
           'politician',
           'suspicion',
           'sabotag',
           'connect',
           'strike',
           'action',
           'presid',
           'slobodan',
           'milosev',
           'listen',
           'bbc',
           'news',
           'world']

In [111]: with open('result.txt', 'w') as f:
              f.write(text[0])
              for element in text[1:]:
                  f.write(element)
                  f.write('\n')
```

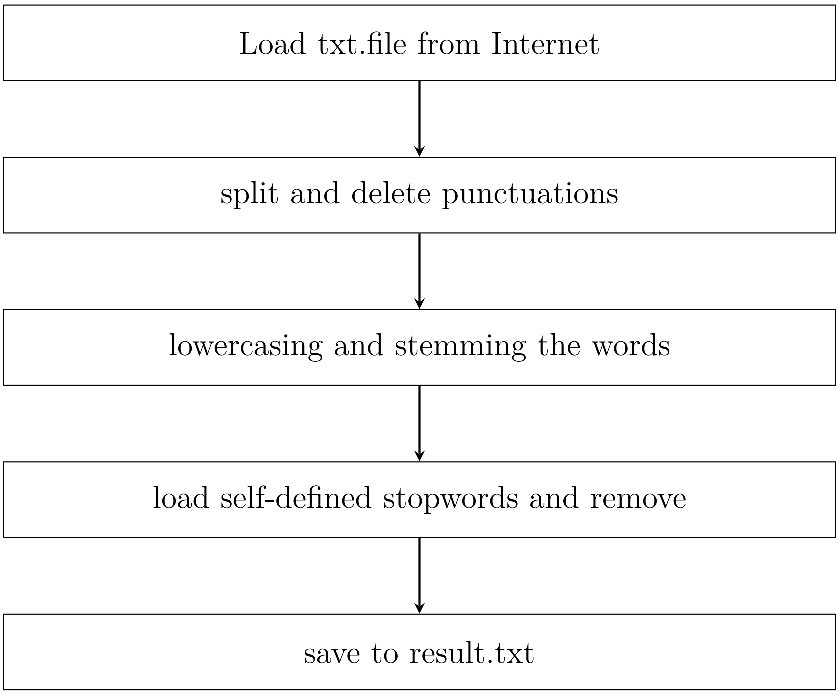
Figure 1: ipynb



```
yugoslavauthor
plan
arrest
eleven
coal
miner
two
opposit
politician
suspicion
sabotag
connect
strike
action
presid
slobodan
milosev
listen
bbc
news
world
```

Figure 2: result.txt

4 Program Logic


Load txt.file from Internet
split and delete punctuations
lowercasing and stemming the words
load self-defined stopwords and remove
save to result.txt

5 Environment