按兩下 (或按 Enter 鍵) 即可編輯

按兩下 (或按 Enter 鍵) 即可編輯

Lab#1, NLP Spring 2023

This is due on 2023/03/06 15:30, commit to your github as a PDF (lab1.pdf) (File>Print>Save as PDF).

IMPORTANT: After copying this notebook to your Google Drive, please paste a link to it below. To get a publicly-accessible link, hit the *Share* button at the top right, then click "Get shareable link" and copy over the result. If you fail to do this, you will receive no credit for this lab!

LINK: paste your link here

https://colab.research.google.com/drive/1D70P_ArOV-RbLnFGyOxoo7PHUhyCFPPq?usp=sharing

---

Student ID:B0928031

Name:鄭茹云

Question 1 (100 points)

Let's switch over to coding! Write some code in this cell to compute the number of unique word **tokens** in this paragraph (5 steps of Text Normalisation: 1. Lowercase Conversion, 2. Remove punctuations, 3. Stemming, 4. Lemmatisation, 5. Stopword Removal). Use a whitespace tokenizer to separate words (i.e., split the string by white space). Be sure that the cell's output is visible in the PDF file you turn in on Github.

按兩下 (或按 Enter 鍵) 即可編輯

儲存成功!

```
I went to Manderley again. It seemed to me
that I was passing through the iron gates that led to the driveway.
The drive was just a narrow track now, its stony surface covered
with grass and weeds. Sometimes, when I thought I had lost it, it
would appear again, beneath a fallen tree or beyond a muddy pool
formed by the winter rains. The trees had thrown out new
low branches which stretched across my way. I came to the house
suddenly, and stood there with my heart beating fast and tears
filling my eyes.'''


# DO NOT MODIFY THE VARIABLES
tokens = 0
word_tokens = []

# YOUR CODE HERE! POPULATE THE tokens and word_tokens VARIABLES WITH THE CORRECT VALUES!
tokens=paragraph.lower()

import nltk
from nltk.tokenize import word_tokenize
nltk.download("punkt")
tlist = nltk.word_tokenize(paragraph)
tlist=[word.lower() for word in tlist if word.isalpha()]
#

from nltk.stem import PorterStemmer,LancasterStemmer,SnowballStemmer

tokens=["the","spectators","all","stood","and","sang","the","national","an","them"]

port=nltk.stem.PorterStemmer()
stemmed_port=[port.stem(token) for token in tlist]

lanc=nltk.stem.LancasterStemmer()
stemmed_lanc=[lanc.stem(token) for token in tlist]

snow=nltk.stem.SnowballStemmer("english")
stemmed_snow=[snow.stem(token) for token in tlist]

from nltk.stem import WordNetLemmatizer
lemmatiser=WordNetLemmatizer()
```

```
nltk.download('wordnet')
nltk.download('omw-1.4')
tlist=[lemmatiser.lemmatize(token)  for  token  in  tlist]

from  nltk.corpus  import  stopwords
nltk.download("stopwords")

stop_words  =  set(nltk.corpus.stopwords.words("english"))
tlist  =  [token  for  token  in  tlist  if  token  not  in  stop_words]

word_tokens  =  tlist
tokens=len(tlist)


#  DO  NOT  MODIFY  THE  BELOW  LINE!
print('Number  of  word  tokens:  %d'  %  (tokens))
print("printing  lists  separated  by  commas")
print(*word_tokens,  sep  =  ",  ")
```

```
Number of word tokens: 53
printing lists separated by commas
last, night, dreamed, went, manderley, seemed, wa, passing, iron, gate, led, driveway, drive, wa, narrow, track, stony, surface, covered, grass,
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```