▾ 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/1VR1WZuLvgpcVtWuvUY08kz7a6KvdGnsP?usp=sharing

---

Student ID:B0928013

Name:吳佳恩

▾ 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

```python
from nltk.stem.snowball import SnowballStemmer
paragraph = '''Last night I dreamed 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!

#小寫
tokens = paragraph.lower()
print("小寫\n", tokens)
#移除標點符號
import nltk
from nltk.tokenize import word_tokenize
nltk.download('punkt')
words = nltk.word_tokenize(paragraph)
print("\n移除標點符號")
words=[word.lower() for word in words if word.isalpha()]
print(words)

#語幹提取
print("\n語幹提取")
from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer
port = PorterStemmer()
stemmed_port = [port.stem(token) for token in words]#
print(stemmed_port)

lanc = LancasterStemmer()
stemmed_lanc = [lanc.stem(token) for token in words]
print(stemmed_lanc)

snow = SnowballStemmer("english")
stemmed_snow = [snow.stem(token) for token in words]
print(stemmed_snow)
#詞型還原
nltk.download('wordnet')
nltk.download('omw-1.4')
print("\n詞型還原")
from nltk.stem import WordNetLemmatizer
lemmatiser = WordNetLemmatizer()
```

```
print(lemmatiser)
lemmatised = [lemmatiser.lemmatize(token) for token in words]#
print(lemmatised)

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

⤷　小寫
　　 last night i dreamed 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.

　　移除標點符號
　　['last', 'night', 'i', 'dreamed', 'i', 'went', 'to', 'manderley', 'again', 'it', 'seemed', 'to', 'me', 'that', 'i', 'was', 'passing', 'through',

　　語幹提取
　　['last', 'night', 'i', 'dream', 'i', 'went', 'to', 'manderley', 'again', 'it', 'seem', 'to', 'me', 'that', 'i', 'wa', 'pass', 'through', 'the',
　　['last', 'night', 'i', 'dream', 'i', 'went', 'to', 'manderley', 'again', 'it', 'seem', 'to', 'me', 'that', 'i', 'was', 'pass', 'through', 'the',
　　['last', 'night', 'i', 'dream', 'i', 'went', 'to', 'manderley', 'again', 'it', 'seem', 'to', 'me', 'that', 'i', 'was', 'pass', 'through', 'the',

　　詞型還原
　　<WordNetLemmatizer>
　　['last', 'night', 'i', 'dreamed', 'i', 'went', 'to', 'manderley', 'again', 'it', 'seemed', 'to', 'me', 'that', 'i', 'wa', 'passing', 'through',

　　Number of word tokens: 100
　　printing lists separated by commas
　　last, night, i, dreamed, i, went, to, manderley, again, it, seemed, to, me, that, i, was, passing, through, the, iron, gates, that, led, to, the
　　[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!

◀ ▬▬▬▬▬▬ ▶

✓ 0 秒　完成時間: 晚上8:34　　　　　　　　　　● ✕