

▼ 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/1LaDTR5Idye2YSOhxbn3AuzxfXehQN9bw?usp=sharing>

Student ID: B0928007

Name: 余明昌

▼ 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

```
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
```

```
# split the words
```

```
import re
```

```
word_tokens = re.split(" |\n", paragraph)
```

```
# Lowercase Conversion
```

```
word_tokens = [w.lower() for w in word_tokens]
```

```
# Stopword removal
```

```
from nltk.corpus import stopwords
```

```
import nltk
```

```
nltk.download("stopwords")
```

```
stop_words = set(stopwords.words("english"))
```

```
word_tokens = [word for word in word_tokens if word not in stop_words]
```

```
# Remove punctuation
```

```
def remove_punctuation(word):
```

```
    return [word for word in word_tokens if word.isalpha()]
```

```
word_tokens = remove_punctuation(word_tokens)
```

```
# Stemming
```

```
from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer
```

```
port = PorterStemmer()
```

```
# word_tokens = [port.stem(token) for token in word_tokens]
```

```
lanc = LancasterStemmer()
```

```
# word_tokens = [lanc.stem(token) for token in word_tokens]
```

```
snow = SnowballStemmer("english")
```

```
word_tokens = [snow.stem(token) for token in word_tokens]
```

```
# Lemmatisation
```

```
from nltk.stem import WordNetLemmatizer
```

```
lemmatiser = WordNetLemmatizer()
```

```
word_tokens = [lemmatiser.lemmatize(token) for token in word_tokens]
```

```
# assign value to tokens
```

```
tokens = len(word_tokens)
```

```
# DO NOT MODIFY THE BELOW LINE!
```

```
print('Number of word tokens: {}'.format(tokens))
```

```
print('Number of word tokens: %d' % (tokens))  
print("printing lists separated by commas")  
print(*word_tokens, sep = ", ")
```

Number of word tokens: 44

printing lists separated by commas

last, night, dream, went, manderley, seem, pas, iron, gate, led, drive, nar

[nltk_data] Downloading package stopwords to /root/nltk_data...

[nltk_data] Package stopwords is already up-to-date!

[Colab 付費產品](#) - [按這裡取消合約](#)

