Lab#1, NLP Spring 2025

This is due on 2025/03/10 16:00, 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/12RhwFKoKnB5cyf9t8NgXtkekd9hv6tVe#scrollTo=SgNZTjrhcHa0

Student ID:B1129014

Name:黎秉鑫

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 VALUES!

import nltk
from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer, WordNetLemmatizer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords, wordnet
nltk.download('punkt', quiet=True)
nltk.download('punkt_tab', quiet=True)
nltk.download('stopwords', quiet=True)
nltk.download('wordnet', quiet=True)

# 1. Lowercase Conversion
paragraph_lower = paragraph.lower()

# 2. Remove punctuations
word_tokens = word_tokenize(paragraph_lower)
def remove_punct(token):
        return [word for word in token if word.isalpha()]
punct_removed = remove_punct(word_tokens)

# print(*punct_removed, sep = ", ") # for debugging

# 3. Stemming
port = PorterStemmer()
stemmed_port = [port.stem(token) for token in punct_removed]
# lanc = LancasterStemmer()
# stemmed_lanc = [lanc.stem(token) for token in stemmed_port]
# snow = SnowballStemmer("english")
# stemmed_snow = [snow.stem(token) for token in stemmed_lanc]
stemmed = stemmed_port

# 4. Lemmatisation
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
lemmatised = [lemmatizer.lemmatize(token) for token in stemmed]
```

```
#lemmatised = [lemmatizer.lemmatize(word, pos=wordnet.VERB) for word in stemmed]

# 5. Stopword Removal
stop_words = set(stopwords.words('english'))
words_no_stop = [word for word in lemmatised if word not in stop_words]

word_tokens = set(words_no_stop)
tokens = len(word_tokens)


# 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: 51
printing lists separated by commas
beat, narrow, would, cover, beneath, pas, weed, fast, went, stretch, appear, came, thrown, stood, stoni, driveway, iron, drive, surfac, wa, some
```