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/12KA2Pt7Kf36Em8vknY42Sjn2FU3Hj8Me?usp=sharing

---

Student ID: B1129028

Name: 陳奕劭

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

```
from enum import unique
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.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer, WordNetLemmatizer

nltk.download('punkt')
nltk.download('punkt_tab')
nltk.download('stopwords')
nltk.download('wordnet')

# Step 1: Lowercase Conversion
paragraph = paragraph.lower()

# Step 2: Tokenization
tokens = word_tokenize(paragraph)

# Step 3: Remove Punctuation
def remove_punctuation(text):
        return [word for word in text if word.isalpha()]

tokens = remove_punctuation(tokens)

# Step 4: Stemming using PorterStemmer
stemmer = SnowballStemmer("english")
stemmed_tokens = [stemmer.stem(token) for token in tokens]

# Step 5: Lemmatization using WordNetLemmatizer
```

```
lemmatizer = WordNetLemmatizer()
lemmatized_tokens = [lemmatizer.lemmatize(token) for token in stemmed_tokens]

# Step 6: Stopword Removal
stop_words = set(stopwords.words("english"))
word_tokens = [word for word in lemmatized_tokens if word not in stop_words]

# Compute number of unique tokens
tokens = len(set(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
last, night, dream, went, manderley, seem, wa, pas, iron, gate, led, driveway, drive, wa, narrow, track, stoni, surfac, cover, grass, weed, some
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Package punkt_tab is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```