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/1KifhqSG0kOeJlwlX7or9rNcYdG2k0BTr?usp=sharing

---

Student ID: B0929038 **Name**: 吳安捷

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 nltk.corpus.reader import wordnet
from nltk.stem.snowball import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import re
import nltk
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 VALUE
def get_word_pos(tag):
    if tag.startswith('J'):
        return wordnet.ADJ
    elif tag.startswith('V'):
        return wordnet.VERB
    elif tag.startswith('N'):
        return wordnet.NOUN
    elif tag.startswith('R'):
        return wordnet.ADV
    else:
        return None
temp=[]
lower_prargraph=paragraph.lower()
remove_puntuation = re.sub(r'[.,"\'-?:!;]','',lower_prargraph)
temp=nltk.word_tokenize(remove_puntuation)
temp=list(set(temp))
tags = nltk.pos_tag(temp)
lemmatiser=WordNetLemmatizer()
lemmatise_sent=[]
for tag in tags:
    pos_tag=get_word_pos(tag[1])or wordnet.NOUN
    lemmatise_sent.append(lemmatiser.lemmatize(tag[0],pos=pos_tag))
stop_words=set(stopwords.words("english"))
words_no_stop=[word for word in lemmatise_sent if word not in stop_words ]
tokens=len(words_no_stop)
word_tokens=words_no_stop

# 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
think, across, drive, house, night, stretch, dream, appear, lead, beneath, branch, iron, fall
```

✓ 0 秒 完成時間：下午3:44