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/xxxxxxxxx

Student ID:

Name:

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!
paragraphs_lower = paragraph.lower()

def remove_punct(token):
    return [word for word in token if word.isalpha()]

paragraphs_sent = remove_punct(paragraphs_lower)


from nltk import word_tokenize
#3
from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer

#stemming
```

```
port = PorterStemmer()
stemmed_port = [port.stem(token) for token in paragraphs_sent]

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

snow = SnowballStemmer("english")
stemmed_snow = [snow.stem(token) for token in paragraphs_sent]

#4
from nltk import word_tokenize
from nltk.stem import WordNetLemmatizer

lemmatiser = WordNetLemmatizer()
lemmatised = [lemmatiser.lemmatize(token) for token in stemmed_snow]

#5
from nltk.corpus import stopwords
nltk.download("stopwords")

stop_words = set(stopwords.word("english"))

words_no_stop = [word for word in lemmatised if word not in stop_words]



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

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
/usr/local/lib/python3.8/dist-packages/nltk/corpus/util.py in __load(self)
     83             try:
---> 84                 root = nltk.data.find(f"{self.subdir}/{zip_name}")
     85             except LookupError:
```

7 frames

```
LookupError: 
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('wordnet')
  
  For more information see: https://www.nltk.org/data.html

  Attempted to load corpora/wordnet.zip/wordnet/

  Searched in:
    - '/root/nltk_data'
    - '/usr/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/local/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/local/lib/nltk_data'
**********************************************************************


During handling of the above exception, another exception occurred:

LookupError                               Traceback (most recent call last)
/usr/local/lib/python3.8/dist-packages/nltk/data.py in find(resource_name, paths)
    581     sep = "*" * 70
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584 
    585 

LookupError: 
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('wordnet')
  
  For more information see: https://www.nltk.org/data.html

  Attempted to load corpora/wordnet

  Searched in:
    - '/root/nltk_data'
```

0 秒 完成時間：下午3:53