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/1LvqyLhX6eiT07APZTsrypvUoQ8JG3FOY?usp=sharing

---

Student ID: B0928026

Name: 洪詩晴

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!

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

#2. Remove punctuations
import nltk as nltk
nltk.download("punkt")
def remove_punct(token):
  return [word for word in token if word.isalpha()]

sent = remove_punct(paragraph)

#3. Stemming

from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer

tokens = ["the", "spectators", "all", "stood", "and", "sang", "the", "national", "anthem"]

#stemming
port = PorterStemmer()
stemmed_port = [port.stem(token) for token in tokens]

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

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

#4. Lemmatisation

from nltk.stem import WordNetLemmatizer
tokens = ["the", "spectators", "all", "stood", "and", "sang", "the", "national", "anthem"]

lemmatiser = WordNetLemmatizer()
```

```
  for token in tokens:
    lemmatised = lemmatiser.lemmatize(token)

  #5. Stopword Removal

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

  # defining stopwords in English
  stop_words = set(stopwords.words("english"))

  # removing stop words
  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 = ", ")
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
---------------------------------------------------------------------------
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:

6 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'
```


0 秒 完成時間：下午3:54