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/1fjvA_J75Cky4C5izTLgP0ufY8oTtH7IS#scrollTo=hpg8eU9wNBci---

Student ID:B0928029

Name:毛謙鎧

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.

＋ コード ＋ テキスト

編集するにはダブルクリックするか Enter キーを押してください

```
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!

# one
tokens_lower = paragraph.lower()

# two
import nltk
nltk.download("puntk")
def remove_punct (tokens_lower) :
    return [word for word in tokens_lower if word.isalpha()]
sent = remove_punct(tokens_lower)
# 3

# importing stemmer classes
from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer
tokens = ["the","ational", "anthem" ]
# stemming
port = PorterStemmer()
stemmed_port = [port.stem (tokens_lower for tokens_lower in tokens)]

# from nltk.stem import WordNetLemmatizer
# tokens = ["the",
# "spectators",
# "all", "stood", "and", "sang",
# "the","national",
# "anthem"]
# lemmatiser = WordNetLemmatizer()
# lemmatised = (lemmatiser.lemmatize(token) for token in tokens]

#  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(tokens_lower)
print(sent)
print(stemmed_port)
print('Number of word tokens: %d' % (tokens))
print("printing lists separated by commas")
print(*word_tokens, sep = ", ")
```

```
[nltk_data] Error loading puntk: Package 'puntk' not found in index
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-17-34b0f995b368> in <module>
     31 # stemming
     32 port = PorterStemmer()
---> 33 stemmed_port = [port.stem (tokens_lower for tokens_lower in tokens)]
     34 
     35 # from nltk.stem import WordNetLemmatizer

/usr/local/lib/python3.8/dist-packages/nltk/stem/porter.py in stem(self, word, to_lowercase)
    656         :param to_lowercase: if `to_lowercase=True` the word always lowercase
    657         """
--> 658         stem = word.lower() if to_lowercase else word
    659 
    660         if self.mode == self.NLTK_EXTENSIONS and word in self.pool:

AttributeError: 'generator' object has no attribute 'lower'
```

SEARCH STACK OVERFLOW