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/15vN702ONpbn1CsKKX91tbux_z2repjKk?usp=sharing

---

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.

Double-click (or enter) to edit

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
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 VA

#Lowercase conversion
paragraph = paragraph.lower()
import nltk
#nltk.download("punkt")

#remove punctuations
def remove_punct(token):
  return [word for word in token if word.isalpha()]
paragraph = nltk.word_tokenize(paragraph)
paragraph = remove_punct(paragraph)

# print(paragraph)
for word in paragraph:
  if word not in word_tokens:
    word_tokens.append(word)
# print(word_tokens)

#stopword removal
from nltk.corpus import stopwords
# nltk.download("stopwords")

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

word_tokens = [word for word in word_tokens if word not in stop_words]

from nltk.stem import PorterStemmer, LancasterStemmer, SnowballStemmer

port = PorterStemmer()
stemmed_port  = [port.stem(token) for token in word_tokens]
# print(stemmed_port)
lanc = LancasterStemmer()
stemmed_lanc = [lanc.stem(token) for token in word_tokens]

snow = SnowballStemmer("english")
stemmed_snow = [snow.stem(token) for token in word_tokens]
#Lemmatisation
from nltk.stem import WordNetLemmatizer
# nltk.download('wordnet')
# nltk.download('omw-1.4')
lemmatiser = WordNetLemmatizer()
word_tokens = [lemmatiser.lemmatize(token) for token in stemmed_port]


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

```
Number of word tokens: 51
printing lists separated by space
last night dream went manderley seem pas iron gate led driveway drive narrow t
```