# Portfolio Assignment 3

## Liam Leece - lcl180002

An assignment to help me better understand NLTK and get a look at a professional-level NLP API.

```
import nltk
nltk.download("stopwords")
nltk.download("wordnet")
nltk.download("punkt")
nltk.download("omw-1.4")
nltk.download('gutenberg')
nltk.download('genesis')
nltk.download('inaugural')
nltk.download('nps_chat')
nltk.download('webtext')
nltk.download('treebank')
from nltk.book import *
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package gutenberg to /root/nltk_data...
[nltk_data]   Package gutenberg is already up-to-date!
[nltk_data] Downloading package genesis to /root/nltk_data...
[nltk_data]   Package genesis is already up-to-date!
[nltk_data] Downloading package inaugural to /root/nltk_data...
[nltk_data]   Package inaugural is already up-to-date!
[nltk_data] Downloading package nps_chat to /root/nltk_data...
[nltk_data]   Package nps_chat is already up-to-date!
[nltk_data] Downloading package webtext to /root/nltk_data...
[nltk_data]   Package webtext is already up-to-date!
[nltk_data] Downloading package treebank to /root/nltk_data...
[nltk_data]   Unzipping corpora/treebank.zip.
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
```

```
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

## Tokens

This shows the first 20 tokens in text1. One notable thing I noticed is that punctuation and parenthesis are counted as tokens.

```
newtext = text1[:20]
print(newtext)
```

```
['[', 'Moby', 'Dick', 'by', 'Herman', 'Melville', '1851', ']', 'ETYMOLOGY', '.', '(', 'S
```

## Concordance

This displays the concordance method which finds words within a text.

```
text1.concordance("sea",79,5)
```

```
Displaying 5 of 455 matches:
 shall slay the dragon that is in the sea ." -- ISAIAH " And what thing soever
 S PLUTARCH ' S MORALS . " The Indian Sea breedeth the most and the biggest fis
cely had we proceeded two days on the sea , when about sunrise a great many Wha
many Whales and other monsters of the sea , appeared . Among the former , one w
 waves on all sides , and beating the sea before him into a foam ." -- TOOKE '
```

## Count Method

The NLTK API and the Python count methods function exactly the same way. The NLT version even uses the original Python function to calculate its answer. They both take a text and return the number of times the chosen value appears within it.

```
print(text1.count("sea"))
print(text2.count("the"))
```

```
433
3861
```

## Tokenizer

This snippet takes a few sentences from the book, Fundamentals Of Database Systems 7th Edition, and tokenizes it into words using NLTK's built in tokenizer.

Source: http://auhd.edu.ye/upfiles/elibrary/Azal2020-01-22-12-28-11-76901.pdf

```
from nltk import word_tokenize
raw_text = "This book introduces the fundamental concepts necessary for designing, using, and
tokens = word_tokenize(raw_text)
print(tokens[:10])
```

```
['This', 'book', 'introduces', 'the', 'fundamental', 'concepts', 'necessary', 'for', 'de
```

## Sentence Tokenizer

This chunk does much the same thing as the above section but instead tokenizes the text into sentences rather than words.

```
from nltk import sent_tokenize
sent = sent_tokenize(raw_text)
sent
```

```
['This book introduces the fundamental concepts necessary for designing, using, and
implementing database systems and database applications.',
 'Our presentation stresses the fundamentals of database modeling and design, the
languages and models provided by the database management systems, and database system
implementation techniques.',
 'The book is meant to be used as a textbook for a one- or two-semester course in
database systems at the junior, senior, or graduate level, and as a reference book.',
 'Our goal is to provide an in-depth and up-to-date presentation of the most important
aspects of database systems and applications, and related technologies.',
 'We assume that readers are familiar with elementary programming and data-structuring
concepts and that they have had some exposure to the basics of computer organization.']
```

## Porter Stemmer

This code will take the raw text tokens and stem the applicable words into shorter or easier to use versions. This process is useful for gathering data on large chunks of text.

```
from nltk import PorterStemmer
ps = PorterStemmer()
stemtext = [ps.stem(x) for x in tokens]
print(stemtext)
```

```
['thi', 'book', 'introduc', 'the', 'fundament', 'concept', 'necessari', 'for', 'design',
```

## Word Net Lemmatizer

This section does much the same as the Porter Stemmer, but uses a few different qualifications for trimming words. The lemmatizer is also useful for the same reasons as the stemmer.

Some words that are different when shortened, in the format stem-lemma, fundament-fundamental use-using databas-database languag-language applic-application

```
from nltk import WordNetLemmatizer
lm = WordNetLemmatizer()
lemtext = [lm.lemmatize(x) for x in tokens]
print(lemtext)
```

```
['This', 'book', 'introduces', 'the', 'fundamental', 'concept', 'necessary', 'for', 'des
```

## Conclusion

Overall, I enjoyed learning the ins and outs of the NLTK library system. I belive the functionality is good, especially the ease of importing and downloading things from the library. The overall code quality is great but could be improved upon by adding new functionalities to repeated code, such as the count method. I believe I can use NLTK in many future projects to help organize any text files I work with or narrow down the amount of searching for text data required.

✓ 0s completed at 12:03 PM ● ×