2. Install

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

```
[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 genesis to /root/nltk_data...
[nltk_data]   Package genesis is already up-to-date!
[nltk_data] Downloading package gutenberg to /root/nltk_data...
[nltk_data]   Unzipping corpora/gutenberg.zip.
[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]   Package treebank is already up-to-date!
True
```

3. Extract the first 20 tokens from text1

- From the documentation I learned that the Text objects have many built in functions that gives information about specific words such as how many times they appear in the text or where they are in the text.
- The tokens() function just returns the private list of tokens from the ContextIndex class.

```
from nltk.book import text1

text1_tokens = [text1.tokens[i] for i in range(0,20)]
print(text1_tokens)
```

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

4. Print concordance for word 'sea' in text1

```
text1.concordance('sea',lines=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 '
```

5. count() method experiment

- count() counts the number of times a word appears in a text. It is similar to the count() method in Python as that counts the number of elements in a list that have the same value specified.

```
print(text1.count('hello'))
print(text1.count('sea'))
print(text1.count('by a'))
```

```
0
433
0
```

6. Use tokenizer to print first 10 tokens of some raw text

Source: https://en.wikipedia.org/wiki/Animal_Crossing

```
from nltk.tokenize import word_tokenize

raw_text = """In the Animal Crossing games, the player assumes the role of a human characte
              Gameplay is open-ended: players have no defined objectives but are instead er
              Animal Crossing games are played in real-time, utilizing the video game conso
              Thus, the passage of time in the game world reflects that in reality, as wel
              Some in-game events, such as holidays or the growth of a tree, occur at certa
tokens = word_tokenize(raw_text)

for i in range(0,10):
  print(tokens[i])
```

```
In
the
Animal
Crossing
games
,
the
player
assumes
the
```

7. Use sentence tokenizer on same raw text as above

```
from nltk.tokenize import sent_tokenize

segments = sent_tokenize(raw_text)

for sentence in segments:
  print(sentence)
```

```
In the Animal Crossing games, the player assumes the role of a human character who mo
Gameplay is open-ended: players have no defined objectives but are instead encourage
Animal Crossing games are played in real-time, utilizing the video game console's int
Thus, the passage of time in the game world reflects that in reality, as well as the
Some in-game events, such as holidays or the growth of a tree, occur at certain times
```

8. Use PorterStemmer to write a list comprehension to stem the text and display the list

```
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()
stems = [stemmer.stem(t) for t in tokens]
print(stems)
```

```
['in', 'the', 'anim', 'cross', 'game', ',', 'the', 'player', 'assum', 'the', 'role',
```

9. Write list comprehension to lemmatize text using WordNetLemmatizer

- anim-Animal
- cross-Crossing
- charact-character
- popul-populated
- anthropomorph-antropomorphic

```
from nltk.stem import WordNetLemmatizer

wlt = WordNetLemmatizer()
lemmas = [wlt.lemmatize(t) for t in tokens]
print(lemmas)
```

```
['In', 'the', 'Animal', 'Crossing', 'game', ',', 'the', 'player', 'assumes', 'the', '
```

10.

- The NLTK library has many functions that help in text analysis and can be used in a wide variety of situations. The NLTK objects are set up in a way that allows the developer to write quick and concise code even when performing complicated tasks. One downside to NLTK, however, is the size of the library and the complexity of downloading many different packages. In the future I may use NLTK in projects that involve sentiment analysis or text classification.