Walkthrough Task

Task 1:

```
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from collections import Counter
import matplotlib.pyplot as plt

# New text for analysis
text = """The cat sat on the mat. The mat was comfortable. Cats are curious creatures.
They often explore their surroundings. A curious cat can find trouble in unexpected places.
The mat became a favorite spot for the cat."""

# Step 1: Sentence Tokenization
sentences = sent_tokenize(text)
print("Sentences:\n", sentences)

# Step 2: Word Tokenization and Stop Words Removal
vocab = {}
stop_words = set(stopwords.words('english'))

for sentence in sentences:
    words = word_tokenize(sentence)
    for word in words:
        word = word.lower()
```

```
        if word not in stop_words and word.isalpha():  # Check if the word is alphabetic
            if word not in vocab:
                vocab[word] = 0
            vocab[word] += 1

print("\nVocabulary with Frequencies:\n", vocab)

# Step 3: Extract Most Common Words
vocab_size = 5
common_vocab = Counter(vocab).most_common(vocab_size)
print("\nMost Common Words:\n", common_vocab)

# Step 4: Create Word to Index Mapping
word_to_index = {word[0]: index + 1 for index, word in enumerate(common_vocab)}
print("\nWord to Index Mapping:\n", word_to_index)

# Step 5: Encode Sentences
encoded_sentences = []
for sentence in sentences:
    encoded_sentence = []
    for word in word_tokenize(sentence):
        word = word.lower()
        encoded_sentence.append(word_to_index.get(word, 0))  # Use 0 for OOV
    encoded_sentences.append(encoded_sentence)

print("\nEncoded Sentences:\n", encoded_sentences)
```

```
# Step 6: Visualization of Most Common Words
labels, values = zip(*common_vocab)
plt.bar(labels, values)
plt.title('Most Common Words in Text')
plt.xlabel('Words')
plt.ylabel('Frequency')
plt.show()
```

Task 2:

```
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk import FreqDist
import numpy as np

# New text for analysis
text = """The sun shines brightly in the sky. The sky is blue and clear.
Birds are singing, and flowers are blooming. It is a beautiful day to be outside.
Nature is full of wonders, and every moment is a gift."""

# Step 1: Sentence Tokenization
sentences = sent_tokenize(text)
print("Sentences:\n", sentences)

# Step 2: Word Tokenization
words = [word.lower() for sentence in sentences for word in word_tokenize(sentence) if word.isalpha()]
```

```
print("\nAll Words:\n", words)

# Step 3: Frequency Distribution
vocab = FreqDist(words)
print("\nFrequency Distribution:\n", vocab)

# Step 4: Access Frequency of a Specific Word
word_to_check = "the"
print(f"\nFrequency of '{word_to_check}':", vocab[word_to_check])

# Step 5: Extract Most Common Words
vocab_size = 5
most_common_vocab = vocab.most_common(vocab_size)
print("\nMost Common Words:\n", most_common_vocab)

# Step 6: Create Word to Index Mapping
word_to_index = {word[0]: index + 1 for index, word in enumerate(most_common_vocab)}
print("\nWord to Index Mapping:\n", word_to_index)

# Step 7: Encode Sentences
encoded_sentences = []
for sentence in sentences:
    encoded_sentence = []
    for word in word_tokenize(sentence):
        word = word.lower()
        encoded_sentence.append(word_to_index.get(word, 0))  # Use 0 for OOV (Out Of Vocabulary)
```

```
    encoded_sentences.append(encoded_sentence)

print("\nEncoded Sentences:\n", encoded_sentences)

# Step 8: Display Original Words with Their Encoded Values
for sentence_index, encoded in enumerate(encoded_sentences):
    print(f"\nOriginal Sentence {sentence_index + 1}: {sentences[sentence_index]}")
    print(f"Encoded Sentence {sentence_index + 1}: {encoded}")
```