Name:Swetha.K

Roll No:235229143

In [3]:
```
import re
from collections import Counter
```

In [4]:
```
import nltk
print(nltk.data.path)
```

```
['C:\\Users\\online.CSCENTER/nltk_data', 'C:\\ProgramData\\Anaconda3\\nltk_da
ta', 'C:\\ProgramData\\Anaconda3\\share\\nltk_data', 'C:\\ProgramData\\Anacon
da3\\lib\\nltk_data', 'C:\\Users\\online.CSCENTER\\AppData\\Roaming\\nltk_dat
a', 'C:\\nltk_data', 'D:\\nltk_data', 'E:\\nltk_data']
```

In [5]:
```
nltk.data.path.append("C:\\Users\\online.CSCENTER/nltk_data")
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\online.CSCENTER\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[5]:
```
True
```

In [6]:
```
def process_novels():
    novels = ["austen-emma.txt", "austen-persuasion.txt", "austen-sense.txt"]

    for novel in novels:
        print(f"\nProcessing {novel}:")

        # A. Open and read the text file
        with open(novel, 'r', encoding='utf-8') as file:
            corpus_text = file.read()

        # B. Build a list of individual sentences
        sentences = nltk.sent_tokenize(corpus_text)

        # C. Print the number of sentences
        print(f"Number of sentences: {len(sentences)}")

        # D. Build a flat tokenized word list and the type list
        words = re.findall(r'\b\w+\b', corpus_text.lower())
        types = set(words)

        # E. Print the token and type counts
        print(f"Token count: {len(words)}")
        print(f"Type count: {len(types)}")

        # F. Build a frequency count dictionary of words
        word_freq = Counter(words)

        # G. Print the top 50 word types and their counts
        print("\nTop 50 word types and their counts:")
        for word, freq in word_freq.most_common(50):
            print(f"{word}: {freq}")

        # Observation: Average sentence length
        avg_sentence_length = sum(len(nltk.word_tokenize(sentence)) for senten
        print(f"\nAverage sentence length: {avg_sentence_length:.2f} words")

# Call the function to process the novels
process_novels()
```

```
Processing austen-emma.txt:
Number of sentences: 7493
Token count: 161983
Type count: 7256

Top 50 word types and their counts:
to: 5239
the: 5201
and: 4896
of: 4291
i: 3178
a: 3129
it: 2528
her: 2469
was: 2398
she: 2340
in: 2188
not: 2140
```

In [ ]:

In [ ]: