Task 1

Write a program to count no. of words in a paragraph. Also find the number of words formed with alphabets

In [1]:

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [2]:

```
# download and load the text6 corpus from NLTK
nltk.download("nps_chat")
text6 = nltk.corpus.nps_chat.words()
```

```
[nltk_data] Downloading package nps_chat to
[nltk_data]     C:\Users\bhavy\AppData\Roaming\nltk_data...
[nltk_data]   Package nps_chat is already up-to-date!
```

In [3]:

```
from nltk.book import *
```

```
*** 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
```

In [4]:

```
import re

def count_words(text):
    word_count = 0
    alphabet_word_count = 0

    # Split the text into individual words
    words = re.findall(r'\b\w+\b', text)

    for word in words:
        # Increment the word count
        word_count += 1

        # Check if the word consists only of alphabets
        if word.isalpha():
            alphabet_word_count += 1

    return word_count, alphabet_word_count

# Example usage with Text 1
text1 = """
This is a sample text. It contains several words, some of which are formed with alphab
There are also words with numbers like "hello123" and special characters like "!@#$%".
to count only the words that consist of alphabets.
"""

total_words, alphabet_words = count_words(text1)

print("Total words:", total_words)
print("Words formed only with alphabets:", alphabet_words)
```

```
Total words: 40
Words formed only with alphabets: 39
```

Task 2

Write a program to find out total no. of unique words in a paragraph. Also find the occurrence of each unique word

In [5]:

```
corpus = gutenberg.words('melville-moby_dick.txt')
text = Text(corpus)
```

In [6]:

```
import re
from collections import Counter

def count_unique_words(paragraph):
    # Split the paragraph into individual words
    words = re.findall(r'\b\w+\b', paragraph)

    # Count the occurrence of each word
    word_counts = Counter(words)

    # Get the total number of unique words
    unique_word_count = len(word_counts)

    return unique_word_count, word_counts

# Example usage

paragraph = text1

total_unique_words, word_occurrences = count_unique_words(paragraph)

print("Total unique words:", total_unique_words)
print("Word occurrences:")
for word, count in word_occurrences.items():
    print(word, ":", count)
```

```
Total unique words: 32
Word occurrences:
This : 1
is : 1
a : 1
sample : 1
text : 1
It : 1
contains : 1
several : 1
words : 3
some : 1
of : 2
which : 1
are : 2
formed : 1
with : 2
alphabets : 2
only : 2
There : 1
also : 1
numbers : 1
like : 2
hello123 : 1
and : 1
special : 1
characters : 1
We : 1
need : 1
to : 1
count : 1
the : 1
that : 1
consist : 1
```