Program5: Perform basic frequency analysis on a text

i. Count the frequency of each word in a given text.

ii. Visualize the word frequencies using a bar chart.

What is Frequency Analysis?

Frequency analysis in Natural Language Processing (NLP) is a fundamental technique to extract insights from text data by identifying the most commonly occurring words.

Why is it useful?

Helps in understanding dominant themes in a text.

Useful for keyword extraction and text summarization.

Forms the basis for more advanced NLP tasks like text classification, sentiment analysis, and topic modeling.

How is it done?

Tokenization: Splitting text into individual words.

Normalization: Lowercasing and removing punctuation/special characters.

Counting Words: Using frequency counts (e.g., Counter from Python's collections module).

Visualization: Using graphs like bar charts or word clouds to represent common words visually.

Step 1: Preprocessing the Text

Before performing frequency analysis, text data needs to be cleaned and processed.

Convert text to lowercase (to avoid counting "NLP" and "nlp" separately).

Remove special characters and punctuation (to count words properly). Tokenize the text (split it into individual words).

Optionally, remove stopwords (common words like "the," "is," "and" that do not add much meaning).

Step 2: Counting Word Frequencies

After preprocessing, we count how many times each word appears.

The Counter() class from the collections module helps efficiently count word occurrences.

We can extract the top N most common words to focus on the most relevant words.

Step 3: Visualizing the Data

Why visualize word frequency? Helps in identifying trends and important words at a glance. Useful in market research, and social media analysis.

Common visualization techniques:

Bar charts: Best for structured data representation.

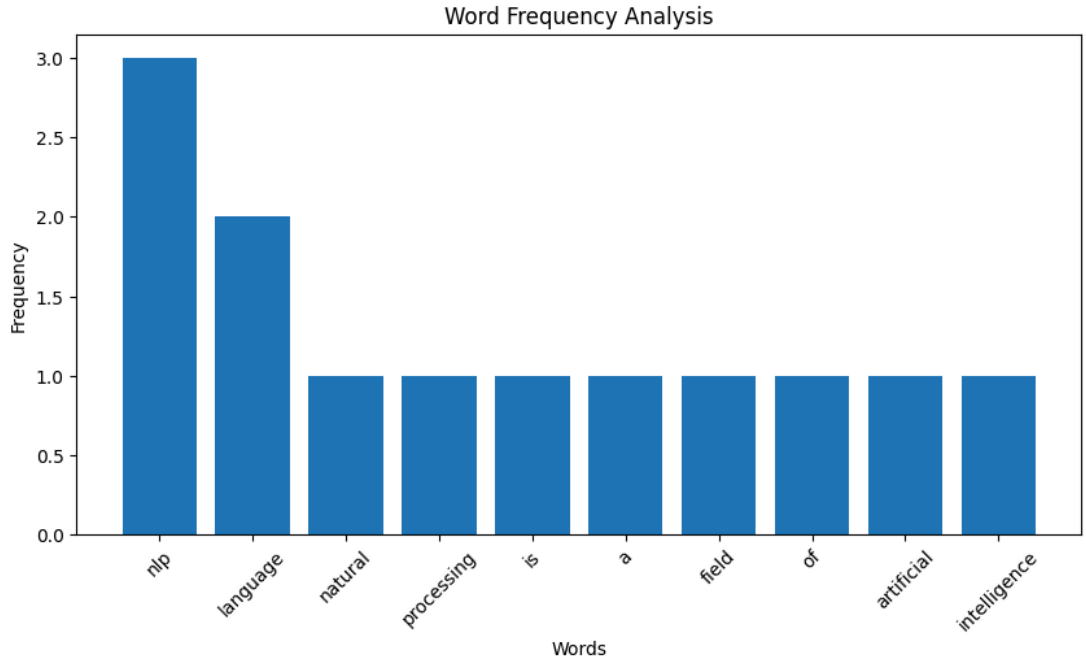
Word clouds: Show the importance of words based on size.

Histograms: Display the distribution of word frequencies.

```
import re
import matplotlib.pyplot as plt
from collections import Counter

# Sample text
text = """Natural Language Processing (NLP) is a field of artificial intelligence.
NLP allows computers to understand human language. NLP techniques include tokenization,
part-of-speech tagging, named entity recognition, and machine translation."""

# Preprocessing: Convert text to lowercase and remove special characters
text = text.lower()
text = re.sub(r'[^a-z\s]', '', text)  # Remove punctuation

# Tokenize words
words = text.split()

# Count word frequencies
word_counts = Counter(words)

# Display top 10 most common words
print("Top 10 words:", word_counts.most_common(10))

# Visualization using Bar Chart
plt.figure(figsize=(10, 5))
```

```
plt.bar(*zip(*word_counts.most_common(10)))  # Plot top 10 words
plt.xlabel("Words")
plt.ylabel("Frequency")
plt.title("Word Frequency Analysis")
plt.xticks(rotation=45)
plt.show()
```

```
Top 10 words: [('nlp', 3), ('language', 2), ('natural', 1), ('processing', 1), ('is', 1), ('a', 1), ('field', 1), ('of', 1), ('artificia
```




```
import re
import nltk
import matplotlib.pyplot as plt
from collections import Counter
from nltk.corpus import stopwords
from wordcloud import WordCloud

# Download stopwords if not already available
nltk.download('stopwords')

# Define stopwords
stop_words = set(stopwords.words('english'))

# Sample text
text = """Natural Language Processing (NLP) is a field of artificial intelligence.
NLP allows computers to understand human language. NLP techniques include tokenization,
part-of-speech tagging, named entity recognition, and machine translation."""

# Preprocessing: Convert to lowercase and remove special characters
text = text.lower()
text = re.sub(r'[^a-z\s]', '', text)  # Remove punctuation

# Tokenization and Stopword Removal
words = text.split()
filtered_words = [word for word in words if word not in stop_words]  # Remove stopwords

# Count word frequencies
word_counts = Counter(filtered_words)

# Display top 10 most common words
print("Top 10 words:", word_counts.most_common(10))

# Bar Chart Visualization
plt.figure(figsize=(10, 5))
plt.bar(*zip(*word_counts.most_common(10)))  # Plot top 10 words
plt.xlabel("Words")
plt.ylabel("Frequency")
plt.title("Word Frequency Analysis (After Stopword Removal)")
plt.xticks(rotation=45)
plt.show()

# Generate Word Cloud
wordcloud = WordCloud(width=800, height=400, background_color='white').generate_from_freq

# Display Word Cloud
plt.figure(figsize=(10, 5))
```

```
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.title("Word Cloud Visualization")
plt.show()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
Top 10 words: [('nlp', 3), ('language', 2), ('natural', 1), ('processing', 1), ('field', 1), ('artificial', 1), ('intelligence', 1), ('a
```