Assignment 6

Objective: Introduce students to basic Natural Language Processing (NLP) techniques, focusing on preprocessing text data and exploring word frequency.

Dataset: Students are required to collect or create a small dataset of text documents (e.g., tweets, news articles, or survey responses). Alternatively, instructors may provide a sample dataset.

Tasks:

1. **Data Loading and Exploration**:
    - Import your text dataset using Python (e.g., pandas).
    - Display basic statistics about the dataset (e.g., number of documents, average document length).
    - Print the first 5 documents in the dataset.
2. **Text Preprocessing**:
    - Perform the following preprocessing steps on the text data:
        - Tokenization
        - Lowercasing
        - Removal of punctuation and special characters
        - Stopword removal (use NLTK or SpaCy stopword lists)
    - Briefly explain each preprocessing step in comments.
3. **Word Frequency Analysis**:
    - Create a **Bag of Words (BoW)** representation of the text using CountVectorizer.
    - Visualize the most frequent words using a **bar chart**.
    - Create and display a **wordcloud** to visualize the frequent words.
4. **TF-IDF Analysis**:
    - Convert the text to a **TF-IDF matrix** using TfidfVectorizer.
    - Briefly explain the significance of TF-IDF in identifying important terms.
5. **Optional (Challenge Task)**:
    - Implement **Topic Modeling Analysis**.
    - Identify and describe 3 dominant topics in your dataset.

Submission:

Submit a Jupyter Notebook containing:

- Code for each task.
- Outputs and visualizations for the tasks.
- Brief comments explaining your findings.

Data Access:

Students can collect their text data from public sources, such as:

- Online Text Repositories: E.g., Kaggle.
- Manual Collection: Gather responses from surveys or web-scraped data (ensure ethical considerations).