Methodology:

Firstly we collected the dataset for preprocessing purposes. The preprocessing steps involve:

- Tokenization: Extracting tokens or words from the dataset.
- Stop words removal: Removing unnecessary words that are called stopwords that offer no help in information retrieval.
- Stemming: Stemming is the process of extracting root words from the dataset used.
- Case conversion: All the words extracted after stemming are converted to lowercase words.

List of necessary packages used:

- NLTK: In this practical, we used the nltk library for the stemming purpose. In that, we used the Porter Stemmer algorithm for stemming purposes.

Dataset used:

The dataset used for this practice is purely handcrafted. I have chosen two genres namely music and the IT sector. In music, I have chosen five bands and the documents are taken from Wikipedia. For the IT sector, I have chosen the FAANG companies and also taken the documents from Wikipedia.

Sample Input:

Documents from the dataset in the form of text files.

Sample output:

Completely preprocessed list of words that occurred in the dataset.

Learning:

From this experiment, we hereby learn how to perform tokenization, stop word removal, and stemming.