Natural Language Processing

Abstract:

Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important.

Problem Statement:

Perform tokenization and POS tagging on the text given reuters data

Dataset Information:

The data set used in this case study is 'reuters' which is a benchmark dataset for document classification. It is a multi-class and multi-label data set having 90 classes, 7769 training documents, and 3019 testing documents. The training set has a vocabulary size of 35247 and can be imported from nltk.corpus

Scope:

- Checking how many characters, words and sentences are there in the corpus
- Removing stopwords from the corpus
- Checking the synonyms for a given word
- Performing tokenization, and POS tagging on reuters news data set using NLTK

Learning Outcome:

The students will learn various aspects of NLP like tokenization, stemming, POS tagging which can be useful to perform sentiment analysis.