Module 18: Natural Language Processing (NLP)

Glossary

Bag of Words

A model that simplifies representations used in NLP; it represents the text as a bag of words, disregarding grammar and even word order while retaining its multiplicity

Corpus

A collection of documents

Lemmatization

A process that analyzes words and returns their base forms

Naive Bayes

A technique for constructing classifiers

NLTK

Natural Language Toolkit — a leading platform for building Python programs to work with human language data

Stemming

A word analysis technique that removes the suffix of a word to derive a base word

Stop Words

Words that are filtered out of the results because they bring no meaning (such as "and")

TF–IDF

Term frequency-inverse document frequency indicates the importance of a particular word to a document in a collection.

Token

A piece of text, such as a word, character, or subword

Vectorization

A feature extraction step that obtains distinct features from the text for model training by converting the text into numerical vectors