AFRICA DATA SCHOOL

Natural language Processing (NLP)

Natural Language Processing or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages.

NLP is used to understand the structure and meaning of human language by analyzing different aspects like syntax and semantics.

Use Cases

- speech recognition,
- document summarization,
- machine translation,
- spam detection,
- named entity recognition,
- question answering,
- autocomplete, predictive typing
- Grammar Correction Tools
- Chatbot
- Smart Assistants

Tools for Natural Language Processing

- CoreNLP from Stanford group
- NLTK, the most widely-mentioned NLP library for Python
- TextBlob, a user-friendly and intuitive NLTK interface
- Gensim, a library for document similarity analysis
- SpaCy, an industrial-strength NLP library built for performance

Basics of different techniques related to Natural Language Processing.

What are Corpus, Tokens, and Engrams?

A **Corpus** is defined as a collection of text documents for example a data set containing news is a corpus or the tweets containing Twitter data is a corpus. So corpus consists of documents, documents comprise paragraphs, paragraphs comprise sentences and sentences comprise further smaller units which are called **Tokens**.

Tokens can be words, phrases, or Engrams, and **Engrams** are defined as the group of n words together.For example, consider this given sentence-“I love my phone.”In this sentence, the uni-grams(n=1) are: I, love, my, phone , Di-grams(n=2) are: I love, love my, my phone

Basics of NLP for Text

- Sentence Tokenization- Sentence tokenization (also called **sentence segmentation**) is the problem of **dividing a string** of written language **into** its component **sentences**.
- Word Tokenization - Word tokenization (also called **word segmentation**) is the problem of **dividing a string** of written language **into** its component **words**.
- Text Lemmatization - refers to the morphological analysis of words, which aims to remove inflectional endings. It helps in returning the base or dictionary form of a word, which is known as the lemma.
- Stemming - Stemming is a kind of normalization for words. It is a technique where a set of words in a sentence are converted into a sequence to shorten its lookup.
- Stop Words - Stop words are words which are **filtered out** before or after processing of text.
- Regex- A **regular expression**, **regex**, or **regexp** is a sequence of characters that define a **search pattern**. Let's see some basics.
- Bag-of-Words - The **bag-of-words** model is a **popular** and **simple feature extraction technique** used when we work with text. convert the text into vectors of numbers.
- TF-IDF - TF-IDF, short for **term frequency-inverse document frequency** is a **statistical measure** used to evaluate the importance of a word to a document in a collection or corpus.

  Part of Speech(PoS) - PoS tags is the properties of words that define their main context, their function, and the usage in a sentence.