



Data Science with Python: Text Preprocessing in NLP #392

By Deepika Pavundoss
Serial No: 161
Batch No: 5

Why Text Pre-processing?

- ❖ A text may contain non-alphanumeric characters, URLs, stop words, words with mixed cases, etc.
- ❖ They sometimes make it difficult for the algorithm to analyse text.
- ❖ To make it easy to analyze different texts and sentences, for Natural Language Processing, the text is always preprocessed.

Steps Involved in Text Pre-processing

- ❖ Removal of punctuations like . , ! ? \$

Eg: (I love having pizzas in evenings!) becomes (I love having pizzas in evenings)

- ❖ Removing Stop words: Removal of the words which do not convey any special meaning and the sentence conveys its message without them.

Eg: I love having pizzas in evenings: (love pizzas evenings) itself conveys the message

Steps Involved in Text Pre-processing

- ❖ Converting text to lower case
- ❖ Lemmatization: to extract the root word without affecting the meaning of the word
Eg: 'enjoyed' becomes 'enjoy'

Many additional steps maybe carried out based on the type data that is preprocessed