Analytics Vidhya

COMMON PRE - PROCESSING TECHNIQUES IN NLP

SWIPE

Analytics Vidhya

Some of the most common pre-processing techniques in NLP include:

- TOKENIZATION
- LOWERCASING
- REMOVE STOP WORDS
- STEMMING / LEMMATIZATION
- VOCABULARY BUILDING
- VECTORIZATION

Analytics Vidhya

1 - TOKENIZATION

Splitting text into individual words or tokens.

Example: NLP is fun

Analytics Vidhya

2 - LOWERCASING

Converting all characters in the text to lowercase to ensure uniformity.

Example: NLP is fun

Analytics Vidhya

3 - REMOVE STOP WORDS

Removing common words that do not carry significant meaning, such as "**and**", "**the**", "**is**" and so on.

Example: NLP is fun

PREV NEXT

Analytics Vidhya

4 - STEMMING

Reducing words to their base or root form.

Example:

"PLAYING" -> "PLAY"

"TEACHING" -> "TEACH"

"FLIES" -> "**FLI**"

"BABIES" -> "**BABI**"

Analytics Vidhya

5 — LEMMATIZATION

Similar to stemming, but it reduces words to their dictionary form (lemma).

Example:

"PLAYING" -> "PLAY"

"TEACHING" -> "TEACH"

"FLIES" -> "**FLY**"

"BABIES" -> "**BABY**"

Analytics Vidhya

5 — VOCABULARY BUILDING

Creating a list of all unique words that appear in a given text dataset.

Example: NLP is fun and I am learning NLP

“nlp”	“is”	“fun”
“I”	“am”	“learning”

*Only the unique words are displayed.

Analytics Vidhya

5 - VECTORIZATION

Process of converting text data into numerical vectors

METHODS OF VECTORIZATION

1. One Hot Encoding
2. Count Vectorizer
3. TF-IDF
4. Word Embedding

Analytics Vidhya

Want to learn with more ways to work with text data?

Join the **FREE** course now!

“Building a Text Classification Model with NLP”

PREV