NLP

Text Preprocessing

- Tokenization
- Stop Words Removal
- Stemming and Lemmatization
- Part of Speech Tagging

Text Preprocessing

→ Text preprocessing is the process of preparing raw text data for further analysis by cleaning and transforming it into a more manageable and consistent format. Common steps in text preprocessing include lowercasing, removing punctuation, and handling special characters.

Example:

Raw Text: "Hello, World! This is an example of Text Preprocessing."

Preprocessed text: "hello world this is an example of text preprocessing"

Tokenization

Tokenization is the process of breaking down text into smaller units called tokens. Tokens can be words, sentences, or subwords. Tokenization is a crucial step as it converts the text into a format that can be easily analyzed.

Example :

Text : "I love NLP!"

Word Tokenization : ["I", "love", "NLP", "!"]

Sentence Tokenization : ['I love NLP.']

Email → Spam / Not Spam

(i) Text Preprocessing → (1) Lower case convert
(2) Punctuation Remove

Text → You won 100000 $

Word Tokenization → ["You", "won", "100000",

Sentences → words

Stop words Removal → common words

Stop words are common words that are often removed from text because they do not carry significant meaning and can clutter the analysis. Examples of stop words include "a," "an," "the," "in," "on," etc

of, to

Example :

Text : " This is a simple example "

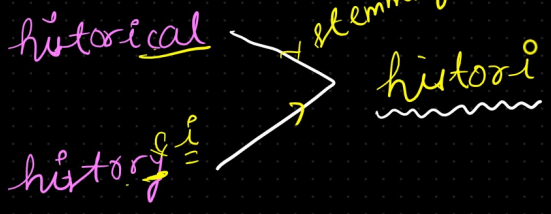
After Stop Word Removal : " This simple example "

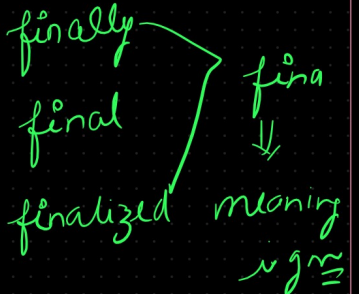
Stemming and Lemmatization

Stemming

Stemming is the process of reducing words to their root form by removing prefixes and suffixes. The resulting root form (stem) may not always be a valid word.

1) Fast
2) Less Extensive





→ It reduces a base word to its stem word. (Root form)

like (Base word) → likes, likely, liking

Use case — Spam classification

Lemmatization

Lemmatization is the process of reducing words to their base or dictionary form (lemma) using morphological analysis. Lemmatization generally produces more accurate results compared to stemming.

(a) Stemming

Studies → suffix — es , stem — studi

studying — suffix — ing , stem - study

lemmization → study

https://www.researchgate.net/publication/348306833_An_Interpretation_of_Lemmatization_and_Stemming_in_Natural_Language_Processing

(1) Meaningful word

Disadvantage

(1) High Computationally Extensive

Use case — Text Summarization, Chatbot

Lang translation

Part-of-Speech Tagging

Part-of-Speech (POS) tagging is the process of assigning a part of speech to each word in a sentence, such as noun, verb, adjective, etc. POS tagging helps in understanding the grammatical structure of the text and is useful for various NLP tasks.

Example:

Text: "The quick brown fox jumps over the lazy dog."

POS Tagging:

The/DT (Determiner)
quick/JJ (Adjective)
brown/JJ (Adjective)
fox/NN (Noun)
jumps/VBZ (Verb)
over/IN (Preposition)
the/DT (Determiner)
lazy/JJ (Adjective)
dog/NN (Noun)