# LECTURE - 3 TEXT PREPROCESSING

## ① Lowercasing

↳ It is a fundamental text preprocessing step that improves the consistency, efficiency, and performance of various natural language processing applications.

## ② Remove HTML Tags

↳ Removal of HTML Tags in text preprocessing is essential to extract and analyze only the textual content from web data, ensuring clean and meaningful text for downstream natural language processing tasks.

## ③ Remove URLs

↳ It is essential to eliminate non-textual elements, ensuring cleaner text data. This enhances the focus on meaningful content and improves the efficiency of natural language processing tasks.

## (4) Remove Punctuations

↳ Removal of punctuations in text preprocessing is crucial to maintain consistency and focus on word semantics. It aids in simplifying text for analysis, avoiding noise, and improving the accuracy of language models.

## ⑤ Chat Word Treatment (sms_slang_translator/slang.txt)

↳ Chat word treatment involves transforming informal or abbreviated words used in online chat into their standard English equivalents, enhancing readability and consistency in text data.

↳ Removing chat word treatment in text processing ensures normalization and standardization, preventing misinterpretation and improving the accuracy of downstream natural language processing tasks by aligning with standard language conventions.

## ⑥ Spelling Correction

↳ Wrong spellings can adversely affect text preprocessing by introducing inconsistencies reducing the effectiveness of tokenization and feature extraction. They hinder accurate analysis, impacting downstream tasks such as sentiment analysis or named ~~activity~~ entity recognition, leading to misinterpretation and errors in NLP models.

## ⑦ Removing Stop Words

↳ Stop words are common words like "His", "is", "and", "of", etc., that are often removed during text preprocessing as they carry minimal semantic meaning and can introduce noise.

↳ Removing stop words in text preprocessing reduces noise, improves computational efficiency, and focuses on more informative words, enhancing the performance of natural language processing tasks.

## ⑧ Handling Emojis

↳ It is crucial for preserving sentiment and meaning. Emojis convey emotions not captured by words alone, impacting sentiment analysis. Proper handling ensures accurate interpretation, preventing loss of valuable information during natural language processing tasks.

## (9) Tokenization

↳ It is the process of breaking a text into individual units called tokens. Tokens can be words, phrases, or even individual characters, depending on the context.

↳ Tokenization is essential in NLP for various tasks such as text analysis, sentiment analysis, and machine learning. It forms the foundation for understanding and processing textual data by converting it into manageable and meaningful ~~comp~~ components, facilitating further analysis and feature extraction.

## (10) Stemming

↳ In grammar, "inflection" is the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood.

Eg: walk → walking, walked, walk!

↳ Mostly used in Information Retrieval System like Google.

NLTK → Library

↓

→ Algorithm ; helps to achieve Stemming.

Stemmer

↓ ↓

Porter Stemmer — for English Language

Snow Ball Stemmer — for Other Languages.

↳ Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.

↳ Stemming is important in text preprocessing for NLP tasks. It reduces words to a common base, enhancing consistency and simplifying text analysis. This process aids in improving the efficiency of text-based applications, information retrieval, and ML models by reducing the dimensionality of the feature space.

## (11) Lemmatization

↳ Lemmatization, unlike stemming reduces the inflected words properly, ensuring that the root word belongs to the language.

↳ In Lemmatization the root word is called a "Lemma". A "Lemma" is the canonical form, dictionary form, or citation form of a set of words.

↳ Slower than stemming, because in Lemmatization the root words are searched inside a dictionary like WordNet which is a Lexical dictionary.