DATA SCIENCE WITH PYTHON : TEXT PREPROCESSING IN NLP

Name: Deepthi M

Serial number: 172

Batch number: 05

Text Preprocessing in NLP

- It is the important step while dealing with NLP.
- This step requires the most time and focus one can contribute.
- Text Preprocessing is considered important as it helps in dealing with used text and keeping back the unused text.

The main and important steps/types in Text preprocessing:

- Tokenization
- Spell Corrections
- Parts of Speech Tagging
- Stop Words
- Singularize / Pluralize
- Language Translation
- Stemming
- Lemmatization

Types of Tokenization:

- Tweet Tokenizer
- MWE Tokenizer
- Regular Expression Tokenizer
- Whitespace Tokenizer
- Word Punkt Tokenizer

Types of Stemming:

- Regexp Stemmer
- Porter Stemmer

Tokenization:
Splitting of the text into smaller tokens/components.

Spell Corrections:
If any word is misspelled, treating that is most important in order to avoid the confusion for the model.

Parts of Speech Tagging:
The text after tokenized, each tokens are given parts of speech labels i.e which parts of speech the token belongs To.

Stop Words:
Stop words are removed from the text as it is considered as non-important component in the model. As the stop Words does not convey the sentiments/emotion in the text.

Singularize / Pluralize:
All the text are either converted to singular or plural so that all the text convey the same count for the model.

Language Translation:
If the given text is not in English then it is converted to English.

Stemming and Lemmatization:
The base words are extracted from the tokens and this process is known as stemming. Lemmatization is the better Version of stemming.

The lion lives in jungle and it is the king of its own wolrd.

The above is the statement given.

First we are tokenizing the given statement.

Tokens:

The, lion, lives, in, jungle, and, it, is, the, king, of, its, own, wolrd

After tokenization, we apply Spell Corrections, Parts of Speech Tagging, Stop Words, Singularize / Pluralize, Stemming, Lemmatization

Spell Correction:

In the above statement, wolrd is mis-spelled, so after spell correction it will retain the correct text.

Wolrd to world.

Stop words removal:

After removing of the stop words the result text is:

The, lion, lives, jungle, king, world

Singularize:

In the above statement all the text is singular, if there might be presence of plural word, then it is singularized.

Stemming:

Extraction of the base words from the preprocessed text.

In the above sentence, live is the base word of lives, hence lives is converted to live.

Advantages of doing the text-preprocessing:

- Transforms the text into more desirable form.
- ML models perform better.
- Improves the efficiency.
- Decreases the training time.

Disadvantages of doing the text-preprocessing:

- The original sentiment of the text might/might not hold.

- Conclusion:

Transforms the text into more desirable form that helps the ML models to perform better which has a impact on the efficiency and also decreases the training time.