what is text processing in NLP

Text processing refers to only the analysis, manipulation, and generation of text, while natural language processing refers to the ability of a computer to understand human language in a valuable way. Basically, natural language processing is the next step after text processing.

Why is Text processing essential?

Text processing is one of the most common tasks used in machine learning applications such as language translation, sentiment analysis, spam filtering, and many others.

Text processing methods

The basis of text processing is mathematics and statistics. You can use all these statistical methods to process and analyze text from a frequency distribution, collocation, concordance, and TF-IDF.

1. Word Frequency

This statistical method accurately determines the most frequently used words or expressions in a particular section of text. With this specific insight, you can address problematic situations, identify areas of success, and more.

2. Collocation

This method helps identify co-occurring words – meaning they commonly occur together. The most frequent kinds of collocations in text are bigrams (two adjacent words) and trigrams (three adjacent words). For example, keeping in touch or launching a product are standard connections.

3. Concordance

concordances effectively help to decipher the ambiguity of human language. The term "problem," for instance, can refer to a number of situations, including an issue, a situation, a topic, or the process of supplying something:

- There was a problem with my account → problem
- We have to solve → situations

- It is a crucial topic → topic
- Your tracking number has been issued → delivered

4. TF-IDF

TF-IDF stands for Inverse Document Frequency. This metric measures how important a word is to a document but is offset by the number of documents that contain the word.

5. Text Summarization

Text summarization is the practice of applying natural language processing to reduce complex technical, scientific, or other jargon to its most straightforward components.

This may seem daunting – our languages are complex. But by using basic algorithms for concatenating nouns and verbs, text summarization software can quickly synthesize complicated language to produce concise output.

6. Text classification

Text classification includes several subdivisions, including topic modeling, sentiment analysis, and keyword extraction.
Text classification takes your text dataset and then structures it for further analysis. It is often used to extract valuable data from customer reviews and customer service logs.

7. Keyword extraction

Keyword extraction is a natural language processing (NLP) technique that identifies and extracts the most important words and phrases from a text.

8. Lemmatization and stemming

Lemmatization and stemming, which is more complex than our other topics, is the segmentation, labeling, and reorganization of textual data according on a root stem or definition.