Text processing:

Text Processing pertains to the analysis of text data using a programming language such as Python. Text Processing is an essential task in NLP as it helps to clean and transform raw data into a suitable format used for analysis or modeling.Unstructured text data can be automatically analyzed and sorted by text processing to obtain useful information.

Text processing methods:

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:

By examining how particular words are employed in various settings, 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.

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.

IDF is calculated as follows where ***t*** is the term (word) we are looking to measure the commonness of and ***N*** is the number of documents (d) in the corpus (D).. The denominator is simply the number of documents in which the term, ***t***, appears in.so tf-idf(t, d) = tf(t, d) * idf(t).

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.

6. Text classification:

Again, text classification organizes large amounts of unstructured text (meaning the raw text data you receive from your customers). Text classification includes several subdivisions, including topic modeling, sentiment analysis, and keyword extraction (which we'll discuss next). 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:

The last key to the text analysis puzzle, keyword extraction, is a broader form of the techniques we've already discussed. The most pertinent information from text is automatically extracted using machine learning and artificial intelligence (AI) techniques.

8. Lemmatization and stemming:

Stemming is the process of getting the root form of a word. Stem or root is the part to which inflectional affixes (-ed, -ize, -de, -s, etc.) are added. The stem of a

word is created by removing the prefix or suffix of a word. So, stemming a word may not result in actual words.Example:looked ---> look,denied ---> deni.

Like stemming, lemmatization also converts a word to its root form. The only difference is that lemmatization ensures that the root word belongs to the language. We will get valid words if we use lemmatization. In NLTK, we use the WordNetLemmatizer to get the lemmas of words. We also need to provide a context for the lemmatization. So, we add the part-of-speech as a parameter.Example:Looked->look,denied->deny.