Date. : / /

Q1.) What are StopWords ?

ANS. - Stopwords are words like, "the", "is", "and" and "in" that are frequently used in a language but hold little semantic value.

. In context of NLP or document processing, they're considered insignificant for the Analysis because they don't help in understanding the core meaning of a text.

. The primary reason to identify and remove stopwords is to improve the Efficiency and Accuracy of text Analysis task.

- By Eliminating these common words (stop words), meaningful terms

. They are useful in Applications like:-

- o Search Engines and Information Retrieval.
- o Text Classification and Topic Modeling.
- o Sentiment Analysis.

- Examples include:-

① Articles : a, an, the
② Conjuction : and, but, or
③ Pro-nouns : He, she, it.
④ Common verbs : is, am, as, were, was, etc.

Q2.) Why is stopword Removal important in text preprocessing?

ANS. - stopword removal is an important step in text pre-processing.

- The vital reasons for the same are:-

1. Reducing Noise and improving Focus.

. stopwords are common words (e.g. 'the', 'is' 'an' 'in') that appear frequently in a language but often carry a very little to no-significant meaning in their own.

. By removing these words, the focus of the analysis shifts to more meaningful and semantically imparts. This helps

Page No.
Date. : / /

to reduce the Noise.

2. Enhancing Efficiency and Performance
- When working with large amounts of textual data, the presence of stopwords can significantly increase the size of a dataset and the computational resource required for processing.

3. Enhancing the Accuracy.
- It can be particularly beneficial for certain NLP tasks such as when searching for information.
- Removing stopwords from a query can lead to a precise search conflict.
- It also helps with text classification models by allowing them to focus on words that are more important.

Q3) How does stopword Removal improve the performance of Machine Learning Models?

ANS.
- Stopword removal is a common pre-processing step in the Natural Language Processing (NLP) that can significantly improve the performance of Machine Learning.

(1.) Reducing Noise and Increasing Focus.
- Stopwords like 'the', 'is', 'and', 'in' are the common words that appear frequently in the text., but they generally don't carry a significant meaning For many NLP tasks.
- By removing the stop words, the Machine learning (model) can focus on important words.

Date. : / /

(2.) Decreasing Data Dimensionality in Text Analysis.

- The total Number of Unique words is the Dimensionality of Data.
- Since, stopwords are the high frequency words, they contribute significantly to this count, creating a large, sparse feature space.
- By removing them we can effectively reduce the number of features that provide:
  - i.) Speed-up training time.
  - ii.) Reduces Memory Usage
  - iii.) Less likely Overfitting.

(3.) Enhancing Model Accuracy.

- For task like IR and Document Clustering, removing stopwords can improve accuracy by allowing model to focus on the most relevant keywords.