Assignment No: 1

Title: Pre-Processing of a text document.

Problem Definition:

Write a program for Pre-processing of a text document such as stop word removal, stemming.

Outcome:

Students will be able to,

1. Apply various tools and techniques for information retrieval and web mining .
2. Evaluate and analyze retrieved information.

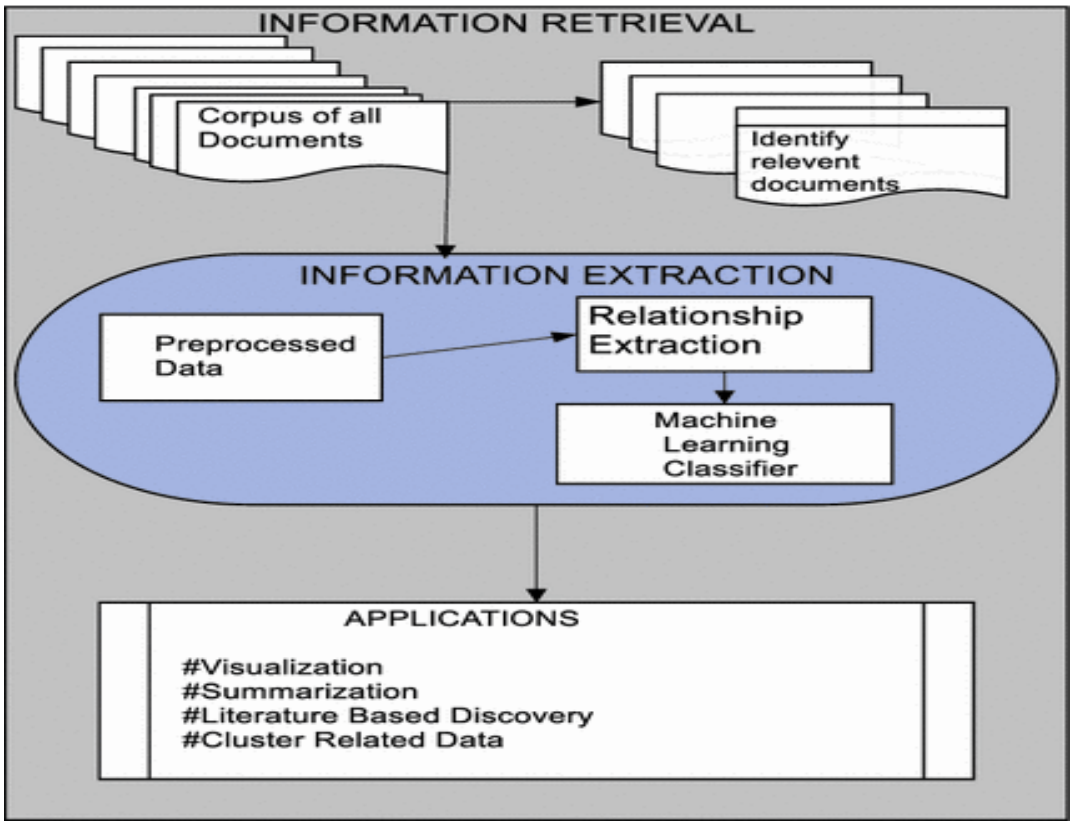
Theory:

Introduction of pre-processing text document:

The text pre-processing process involves unitization and tokenization, standardization and cleaning, stop word removal, and lemmatization or stemming. A custom stop word dictionary can be created to eliminate noise in the text.

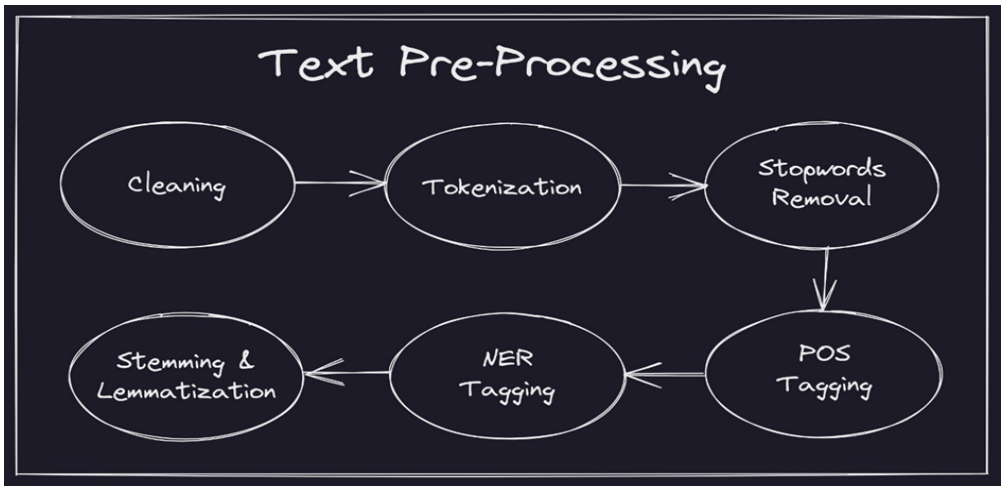
Natural Language Processing (NLP) is a branch of Data Science which deals with Text data. Apart from numerical data, Text data is available to a great extent which is used to analyze and solve business problems. But before using the data for analysis or prediction, processing the data is important.text document pre-processing in information retrieval is a crucial step that involves cleaning, transforming, and enhancing the raw text data to improve the effectiveness and efficiency of retrieval systems. The primary goal of text document pre-processing is to

convert unstructured text into a structured and normalized representation, making it easier to index and retrieve relevant information.



The information retrieval is the task of obtaining relevant information from a large collection of databases. Pre-processing plays an important role in information retrieval to extract the relevant information. In this paper, a text pre-processing approach text pre-processing for information retrieval (TPIR) is proposed. The proposed approach works in two steps. Firstly, spell check utility is used for enhancing stemming and secondly, synonyms of similar tokens are combined. In this paper, proposed technique is applied to a case study on International Monetary Fund. The experimental results prove the efficiency of the proposed approach in terms of complexity, time and performance.

perform these pre-processing steps using Python and the NLTK library:



1. Stop word removal :

Stop word removal is a crucial step in text preprocessing that involves eliminating common words, known as stop words, from a text document. Stop words are words that appear frequently in a language but often don't carry significant meaning in the context of natural language processing and text analysis. Examples of stop words in English include "the," "is," "and," "in," "of," and "on."

The algorithm is implemented as below given steps.

Step 1: The target document text is tokenized and individual words are stored in array.

Step 2: A single stop word is read from stopword list.

Step 3: The stop word is compared to target text in form of array using sequential search technique.

Step 4: If it matches , the word in array is removed , and the comparison is continued till length of array.

Step 5: After removal of stopword completely, another stopword is read from stopword list and again algorithm follows step 2. The algorithm runs continuously until all the stopwords are compared.

Step 6: Resultant text devoid of stopwords is displayed, also required statistics like stopword removed, no. of stopwords removed from target text, total count of words in target text, count of words in resultant text, individual stop word count found in target text is displayed.

The main purpose of stop word removal is to:

Reduce Noise: By removing stop words, you can reduce the noise in the text data, making it easier to focus on the more meaningful words.

Save Storage Space: Stop words often appear in large quantities and can consume unnecessary storage space in databases or text corpora.

Improve Processing Efficiency: Removing stop words can speed up text processing and analysis because there are fewer words to consider.

2. Stemming:

Stemming in information retrieval is a technique used to reduce words to their root or base form (referred to as the "stem") to improve the effectiveness of search and retrieval systems. It involves the application of linguistic rules or algorithms to remove affixes and variations from words. Stemming is particularly useful for grouping words with similar meanings together and increasing the likelihood of retrieving relevant documents in an information retrieval system.

Applications of stemming :

1. Stemming is used in information retrieval systems like search engines. It is used to determine domain vocabularies in domain analysis.
2. To display search results by indexing while documents are evolving into numbers and to map documents to common subjects by stemming.
3. Sentiment Analysis, which examines reviews and comments made by different users about anything, is frequently used for product analysis, such as for online retail stores.

Conclusion:-

we have to implemented pre-processing of text documents is a fundamental and critical step in information retrieval. It plays a pivotal role in enhancing the accuracy, efficiency, and relevance of search and retrieval systems.

Performance & Understanding	Innovation	Timely Completion	Total	Sign & Date
3	1	1	5	