EXPERIMENT NO 1

AIM: To study and implement Preprocessing of text (Tokenization, Filtration, Script Validation, Stop Word Removal, Stemming)

THEORY:

1. Tokenization:

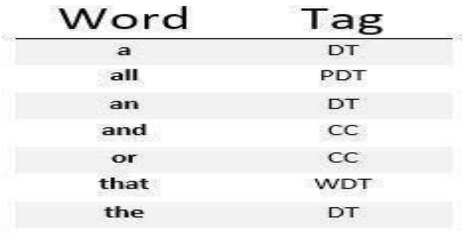
Tokenization is a common task in Natural Language Processing (NLP). It’s a fundamental step in both traditional NLP methods like Count Vectorizer and Advanced Deep Learning-based architectures like Transformers. Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or sub words. Hence, tokenization can be broadly classified into 3 types – word, character, and sub word (n-gram characters) tokenization.

For example, consider the sentence: “Never give up”.

The most common way of forming tokens is based on space. Assuming space as a delimiter, the tokenization of the sentence results in 3 tokens – **Never-give-up**. As each token is a word, it becomes an example of Word tokenization.

2. Filtration:

Many of the words used in the phrase are insignificant and hold no meaning. For example – English is a subject. Here, ‘English’ and ‘subject’ are the most significant words and ‘is’, ‘a’ are almost useless. English subject and subject English holds the same meaning even if we remove the insignificant words – (‘is’, ‘a’). Using the nltk, we can remove the insignificant words by looking at their part-of-speech tags. For that we have to decide which Part-Of-Speech tags are significant.



Word	Tag
a	DT
all	PDT
an	DT
and	CC
or	CC
that	WDT
the	DT

3. Stop Word Removal:

All of the words in a query are stop words. If all the query terms are removed during stop word processing, then the result set is empty. To ensure that search results are returned, stop word removal is disabled when all of the query terms are stop words. For example, if the word *car* is a stop word and you search for *car*, then the search results contain documentsthat match the word *car*. If you search for *car buick*, the search results contain only documents that match the word *buick*.

The word in a query is preceded by the plus
sign (+).The word is part of an exact match.
The word is inside a phrase, for example, "I love my car".

4. Stemming:

Stemming is a part of linguistic studies in morphology and artificial intelligence (AI) information retrieval and extraction. Stemming and AI knowledge extract meaningful information from vast sources like big data or the Internet since additional forms of a word related to a subject may need to be searched to get the best results. Stemming is also a part of queries and Internet engines. Recognizing, searching and retrieving more forms of words returns more results.

Lab Experiments to be Performed in This Session: -

Perform Following Preprocessing Techniques on the given corpus

1. Tokenization,
2. Converting Text Lower Case
3. Remove Numbers
4. Converting Number to Words
5. Remove Punctuation
6. Remove Whitspaces
7. Remove StopWords
8. Count Word Frequency
9. **Stemming (Porter Stemmer and Lancaster Stemmer)**
10. **Lemmatization**