**GROUP 2 (CCS 416)**

1. What is document preprocessing as used in IR

In information retrieval (IR), document preprocessing refers to the set of techniques and operations applied to text documents before they are indexed and used for retrieval purposes. The goal of document preprocessing is to transform the raw text into a more structured and manageable representation that facilitates efficient and effective retrieval of relevant information.

1. Using a clear examples ina each case, differentiate among the following 5 document preprocessing operations
2. Lexical analysis

Lexical analysis involves breaking down the document into individual tokens or words. This step is typically performed through tokenization. For example, given the sentence "I love to eat pizza," the lexical analysis would produce the tokens ["I", "love", "to", "eat", "pizza"].

1. Elimination of stop words

Stop words are commonly used words that do not carry significant meaning and are often removed during preprocessing. For instance, if we have the sentence "The cat is sleeping," the elimination of stop words would result in the tokens ["cat", "sleeping"].

1. Stemming

Stemming is the process of reducing words to their base or root form. It involves removing affixes to obtain the core meaning of a word. For example, stemming the words "running," "runs," and "ran" would result in the common stem "run".

1. Selection of index terms

Index terms are the terms or words that are chosen to represent the content of a document for indexing and retrieval purposes. The selection of index terms involves choosing the most relevant and informative terms that capture the essence of the document. For instance, in the sentence "The big brown dog chased the small black cat," the selected index terms might be ["big", "brown", "dog", "chased", "small", "black", "cat"].

1. Construction of term categorization

Term categorization involves organizing terms into predefined categories or classes based on their semantic or syntactic properties. This can help in enhancing the retrieval process by grouping similar terms together. For example, in a document about animals, the terms "cat," "dog," "lion," and "tiger" might be categorized under the class "mammals."