Q1.> Explain the Basic concept of Text Analysis ?

ANS. Text Analysis ~~a~~ is a process of Analyzing and understanding the written or spoken language. It employs computer Algorithms and techniques to extract valuable information, pattern, and insight From extensive textual data

- In simple terms, text Analytics empowers computers to understand and interpret Human Language.
- Text Analytics has become a crucial tool in todays information Age for two main reasons.

→ i.) Massive growth of Text Data.

→ ii.) Extract valuable insights hidden within data.

- Text Analytics is a powerful tool that finds the meaning and value hidden within mountains of the text data.
- Text Analysis process typically involves/includes the several key steps such as:-

— Identification

— Tokenization.

— Sentence Breaking.

— Part_of_Speech tagging.

— Chunking

— Syntax Parsing.

— Sentence Chaining.

— Keyword Extraction.

— Entity Recognizition.

Q2.> Explain Inverse Document Frequency in Detail.

ANS. Inverse Document Frequency (IDF) is a key-concept in the terms Frequency Inverse - Document Frequency (IF-IDF) weighting scheme, which is commonly used in text Analytics and information veterival.

- IDF measures how important a word is across a set of documents. It down-weights common words and highlights rare, informative ones.

D **FORMULA :-**

The standard Formula For calculating the Inverse Document Frequency (IDF) of term t is:-

$$\boxed{IDF(t) = \log\left(\frac{N}{df_t}\right)}$$

where,

$\boxed{N}$ → Total No. of Documents in the corpus.

$\boxed{df_t}$ → Number of Documents that contain the term t.

- If the IDF of the term is high, it is considered to be more informative.

D **APPLICATION'S:-**

① It is used to search-Engines to rank documents by Relevance.

② It is used for text classification, feature selection and vectorization.

Q3.) Perform Stemming for text = " studies studying cries cry". Compare the results generated with lemmatization Comment on your answer how stemming and lemmatization differ from each other.

ANS. Given text : " studies stydying cries cry",

① The output of stemming is:-

studies ⇒ studi
stydying ⇒ study
cries ⇒ cri
cry ⇒ cry

② Lemmatization:- Output:-

studies ⇒ study

- studying ⇒ study
- cries ⇒ cry
- criy → cry

<u>Compare</u> : stemming and Lemmatization:-

- Stemming chops the words to its root form
  - It uses simple rules and often leads to non-dictionary forms of words like.

    studies → $\boxed{\text{studi}}$ } ...(non - dictionary
    cries – $\boxed{\text{cri}}$ }   root Form)

  - It is a Faster techenique but less acurrate.
- Lemmatization uses vocabulary and Morphological Analysis, returning valid dictonary words like.

    Studies → study } ...(noun root forms)
    cries → cry }

  - It gives more acurate results.

Q4.) Write a Python code for removing stop words from the below documents, convert the document into lowercase and calculate the TF, IDF and TFIDF score for each document.

document A = " Jupiter is the Largest Planet "
document B = " Mars is the Fourth planet From the Sun."

Ans. <u>Python Code:-</u>

```
import nltk
From sklearn.feature_extrurtion.text import TfidFVectorizer
From nltk.corpus
import stopwords.

nltk.download ('stopwords')
doc_a = doc_a.lower()
doc_b = doc_b.lower()
```

```
doc_a =" Jupiter is the Largest Planet"
doc_b =" 'Mars is the Fourth Planet from the Sun!"
doc_a = doc_a.lower()
doc_b = doc_b.lower()
stopwords = set (stopwords.words ('English'))

def remove_stopwords(text):
    return ' '.join([word for word in text.split() if
                     word not in stop_words])


clean_doc_a = remove_stopwords (doc_a)
clean_doc_b = remove_stopwords(doc_b)
vectorizer = TFidFVectorizer()
tfidf_matrix = vectorizer.Fit_transform([clean_doc_a,clean_doc
                                                          -b])
words = vectorizer.get_Feature_names_out()

for i, doc in enumerate([clean_doc_a, clean_doc_b])
   printF(f"\nTF-IDF For Document {'A' if i==0 else 'B'}:"
   for j, word in enumerate(words):
      print(f"{word}: {tfidf_matrix[i,j]:.4f}")
```

OUTPUT:-

```
TF-IDF For Document A:-      | TF-IDF for Document B:-
jupiter : 0.7071             | jupiter  : 0.0000
largest : 0.7071             | largest  : 0.0000
Mars : 0.0000                | mars     : 0.5
Fourth : 0.0000              | fourth   : 0.5
planet : 0.0000              | planet   : 0.5
Sun : 0.0000                 | sun      : 0.5
```