MES Wadia College of Engineering Pune-01

Department of Computer Engineering

Name of Student:	**Class:**
Semester/Year:	**Roll No:**
Date of Performance:	**Date of Submission:**
Examined By:	**Experiment No: Part A-07**

PART: A) ASSIGNMENT NO: 07

Title: Text Analytics

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.

2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.

OBJECTIVES:

- Students should be able to perform Text Analysis using TF IDF Algorithm.

PREREQUISITE:

- Basic of Python Programming.
- Basic of English language.

APPRATUS:

- Programming Language: Python.

ALGORITHM STEP:

Algorithm for Tokenization, POS Tagging, stops words removal, Stemming and Lemmatization:

Step 1: Download the required packages

Step 2: Initialize the text

Step 3: Perform Tokenization

Step 4: Removing Punctuations and Stop Word

Step 5 : Perform Stemming

Step 6: Perform Lemmatization

Step 7: Apply POS Tagging to text

Algorithm for Create representation of document by calculating TFIDF

Step 1: Import the necessary libraries.

Step 2: Initialize the Documents.

Step 3: Create BagofWords (BoW) for Document A and B.

Step 4: Create Collection of Unique words from Document A and B.

Step 5: Create a dictionary of words and their occurrence for each document in the corpus

Step 6: Compute the term frequency for each of our documents.

Step 7: Compute the term Inverse Document Frequency.

Step 8: Compute the term TF/IDF for all words.

CONCLUSION:

QUESTIONS:

1. Explain basic concepts of Text Analytics.
2. Explain Inverse Document Frequency in details.
3. Perform Stemming for *text = "studies studying cries cry*". Compare the results generated wit Lemmatization. Comment on your answer how Stemming and Lemmatization differ from each other.
4. Write Python code for removing stop words from the below documents, convert the documents 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'