



Pune District Education Association's  
**College Of Engineering**

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



DSBDAL

Assignment No:- 07

\* Text Analytics:

1) Extract sample document and apply follows:-

- a) Tokenization
- b) pos Tagging
- c) Stop Words removal
- d) Stemming
- e) lemmatization

2) create representation of document by the calculating term frequency and inverse document frequency.

step 1:-

Extract sample document i.e. Extracting text from doc file.

Here we will extract text from the doc file using docx module.

A) Tokenization:-

1) Tokenization with NLTK.

- NLTK stands for Natural Language Toolkit.
- This is suite of libraries and the programs for statistical natural language processing for english written in python.



## B) Pos Tagging :-

- pos Tagging is a process to mark up the words in text format for a particular part of speech based on its definition and context.
- It is responsible for text reading in a language and assigning some specific token to each other word.
- It also called grammatical tagging.

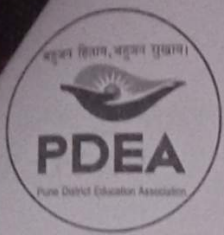
- Steps Involved in the pos tagging example:-

- 1) Tokenize text (Word-tokenize)
- 2) apply pos-tag to above step that is nltk - pos tag (tokenize-text).

NLTK pos Tags Examples are as below:-

Abbreviation	meaning
CC	coordinating conjunction
CD	cardinal digit
DT	determiner
EX	existential there
FW	foreign word
IN	preposition conjunction
IJ	This NLTK pos tag
JJR	adjective





Pune District Education Association's  
**College Of Engineering**

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



c) Stop words removal:-

- Removing stop words with NLTK in python the process of converting data to something a computer can understand is referred to as pre-processing.
- NLTK supports stop word removal, & you can find the list of stop words in the corpus module.
- To remove stop words from a sentence you can divide your text.

d) stemming:-

- stemming words with NLTK.
- stemming is the process of producing morphological variants of a root word.
- stemming programs are commonly referred to as the stemming algorithm.
- ex:- 1) root word = like  
stemming words = "likes"  
"iked"  
"likely"  
"liking"

e) lemmatization:-

- lemmatization is the process of grouping



together the different inflected forms of a word.

-lemmatization is similar to stemming but it brings context to the words, so it links words with similar meaning to one word.

-Applications of lemmatization are -

- 1) used in comprehensive
- 2) used in compact indexing

Example of lemmatization:-

→ rocks : rock

→ corpora : corpus

→ better : good

TF - IDF:

This technique is used to find the meanings of sentences consisting of words and cancels out the incapacities of Bag of Words.

\* formula:-

$tf(t, d)$  : count of  $t$  in  $d$  / number of words in  $d$ .

\* conclusion:-

IDF is the inverse of the document frequency which measures the informativeness of term  $t$ .





# Pune District Education Association's College Of Engineering

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



\*

csv file / dataset:-

Required libraries

```
import nltk
```

```
nltk.download("punkt")
```

```
nltk.download("stopwords")
```

```
nltk.download("wordnet")
```

```
nltk.download("averaged-perceptron-  
tagger")
```

Tokenization:-

बहुजन हिताय, बहुजन सुखाय।

```
from nltk import word_tokenize, sent-  
tokenization  
from corpus
```

pos tagging:-

```
from nltk import pos_tag  
tokens = word_tokenize
```

PDEA Pune District Education Association

stop word removal

```
from nltk.corpus import stopwords  
stop_words = set()
```

stemming

```
from nltk.stem import PorterStemmer  
stemmer = PorterStemmer()
```

lemmatization -

```
from nltk.stem import WordNetLemmatizer
```



TF-IDF -

from sklearn.feature\_extraction.text  
import TfidfVectorizer  
vectorizer = TfidfVectorizer()

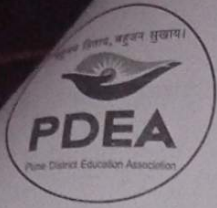
Q.1) list the document preprocessing methods  
→ 1) tokenization 2) pos tagging  
3) stop words removal  
4) stemming  
5) lemmatization.

Q.2) What is mean by pos?  
→ pos means part of speech. It is a process to mark-up the words in text format for a particular part of a speech based on its definition and context also called grammatical tagging.

Eg:- Input :- Everything to permit us

output : ( "Everything", NN ) ( to , to ) ( 'permit  
VB ) , ( 'us' , PRP )

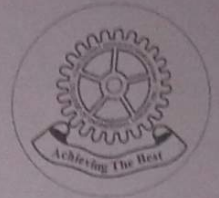
Q.3) What are stop words?  
→ A stop word is a commonly used word (such as "the", "a", "an", "in")



Pune District Education Association's  
**College Of Engineering**

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



Q.4) Explain stemming  
→ Stemming is the process of producing morphological variants of a root/base word.  
e.g. chocolates, chocolates, choco to the root word "chocolate"

Q.5) Explain lemmatization.  
→ - lemmatization is similar to stemming but it brings context to the words.  
- lemmatization does morphological analysis of the words.  
eg. rocks → rock  
corpora → corpus  
better → good

PDEA  
Pune District Education Association