DSBDAL

Assignment No:- 07

* Text Analytics:

1) Extract sample document and apply follows:-

a) Tokenization

b) Pos Togging

c) Stop Words removal

d) Stemming

e) lemmatization

2) create representation of document by the calcuting 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:-

I) 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. tie

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 definanation and context.
- It is responsible for text reading in a language and assiging 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).

NLKT pos Tags Examples are as below-

Abbrevation	Meaning
CC	coordinating conFjunction
CD	cardinal digit
DT	determiner
EX	existential there
FW	foreign Word
IN	preposition conjunction
JJ	This NLTK pos Tag
JJR	adjective

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 reffered to as the stemming algorithm.
- ex:- 1) root word = like
  stemming words = "likes"
  "liked"
  "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 consistening of words and cancels out the incapabilities 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 bocument frequency which measures the informativenesses of term t.

* csv file / dataset:-
Required libraries
import nltk
nltk. download ("punkt")
nltk. download ("stopwords")
nltk. download ("wordnet")
nltk. download ("averaged-perceptron-tagger")

Tokenization:-
from nltk import word-tokenize, sent-tokenization from corpus

Pos tagging:-
from nltk import pos-tag
tokens = word-tokenize

stop word removal
from nltk. corpus import stopwords
stop_words = set()

stemming
From nltk. stem import portestremmer
stemmer = poster stemmer()

lemmatization -
from nltk. stem import wordNetlemmatize

TF - IDF -

```
From Sklearn . feature _ extraction . text
· import TFidvector orizer.
vectorlzer = TFidfvectorrizer()
```

Q.1) list the document preprocessing methods

→ 1) tokenization ② pos tagging
3) stop Words Remoual
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")

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.