SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 008
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – 16AD601 – Natural Language Processing

III Year / VI Semester

Unit 1 – Introduction

Topic 9- Introduction to NLTK

Introduction to NLTK

Python and the Natural Language Toolkit (NLTK)

The Python programing language provides a wide range of tools and libraries for attacking specific NLP tasks.

Many of these are found in the Natural Language Toolkit, or NLTK, an open source collection of libraries, programs, and education resources for building NLP programs.

Statistical NLP, machine learning, and deep learning

The earliest NLP applications were hand-coded, rules-based systems that could perform certain NLP tasks, but couldn't easily scale to accommodate a seemingly endless stream of exceptions or the increasing volumes of text and voice data.

Introduction to NLTK

NLTK (Natural Language Toolkit) is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets.

A variety of tasks can be performed using NLTK such as tokenizing, parse tree visualization etc.

NLTK is a standard python library with prebuilt functions and utilities for the ease of use and implementation.

It is one of the most used libraries for natural language processing and computational linguistics.

Introduction to NLTK

By using NLTK various text processing can be done which include the following

- Tokenization
- Lower case conversion
- Stop Words removal
- Stemming
- Lemmatization
- Parse tree or Syntax Tree generation
- POS Tagging

Introduction to NLTK

Tokenization

Tokenization is the process of breaking text up into smaller chunks as per our requirements.

Word tokenization is the process of breaking a sentence into words. word_tokenize function has been used, which returns a list of words as output.

Sentence tokenization is the process of breaking a corpus into sentence level tokens. It’s essentially used when the corps consists of multiple paragraphs. Each paragraph is broken down into sentences.

Stop Words Removal

Stop words are words which occur frequently in a corpus. e.g a, an, the, in. Frequently occurring words are removed from the corpus for the sake of text-normalization.

Introduction to NLTK

Stemming

It is reduction of inflection from words. Words with same origin will get reduced to a form which may or may not be a word.

In our text we may find many words like playing, played, playfully, etc… which have a root word, play all of these convey the same meaning.

So we can just extract the root word and remove the rest. Here the root word formed is called ‘stem’ and it is not necessarily that stem needs to exist and have a meaning.

Just by committing the suffix and prefix, we generate the stems.

NLTK provides us with PorterStemmer LancasterStemmer and SnowballStemmer packages

Introduction to NLTK

Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.

Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word.

POS Tagging:

Part of Speech tagging is used in text processing to avoid confusion between two same words that have different meanings.

With respect to the definition and context, we give each word a particular tag and process them. Two Steps are used here:

Tokenize text (word_tokenize).

Apply the pos_tag from NLTK to the above step.

THANK YOU