School of Computer Science Engineering and Technology

Course- B. Tech
Course Code- CSET346

Year- 2022
Date: 16-08-2022

Type- Elective
Course Name: Natural language processing
Semester- odd
Batch- ALL

Lab Assignment 01 – First step towards data pre-processing

The main objective of this assignment is to know pre-processing of text data through different steps using natural language tool kit(nltk).

You will learn how to achieve data cleaning, tokenization, stemming, lemmatization and removal of stop words by incorporating importing nltk and its modules.

Perform data cleaning

Perform tokenization for the following sentence.

"""Founded in 2002, SpaceX's mission is to enable humans to become a spacefaring civilization and a multi-planet

species by building a self-sustaining city on Mars. In 2008, SpaceX's Falcon 1 became the first privately developed

liquid-fuel launch vehicle to orbit the Earth.**"""**

Stemming

• A process of removing and replacing suffixes to get to the root form of the word, which is called the stem

• Usually refers to heuristics that chop off suffixes

Lemmatization

• Usually refers to doing things properly with the use of a vocabulary and morphological analysis

• Returns the base or dictionary form of a word, which is known as the lemma

Analyse the utility of the following module-

nltk.stem.PorterStemmer

nltk.stem.WordNetLemmatizer

School of Computer Science Engineering and Technology

Stopwords:

Stop words are a set of commonly used words in any language.

Use the following module for solving the following example

nltk.corpus import stopwords

nltk.download('stopwords')

Analyse the utility of stopwords for the following example

1. This is a sample sentence, showing off the stop words filtration.
2. Nick likes to play footbNick likes to play football, however he is not too fond of tennis.all, however he is not too fond of tennis.