NLP with NLTK in Python (There r other packages also-spacy,skikit)

1. Install Anaconda (in built python 3)
2. Use Anaconda navigator(GUI based) or anaconda command prompt.
3. In Anaconda Navigator, use JypiterNotebook and launch it. Create new notebook by pressing new button and choose python 3
4. If to use anaconda command prompt, click start 🡪anaconda command prompt. It will show

<base>c:\user\...>

Type Python to start

<base>c:\user\...>python

>>> will occur(i.e. command prompt of python)

If u want to exit command prompt, use **ctrl z** and press **enter** key.

1. import nltk-
2. nltk.download()-The NLTK corpora and various modules can be installed by using the common NLTK downloader in the Python interactive shell or a Jupyter Notebook.

Install models,corpus from the window appears

1brown Brown text corpus

2 gutenberg Gutenberg text corpus

3 max\_ne\_chunker Module for text chunking

4 movie\_reviews Movie review sentiment polarity data

5 product\_reviews\_1 Basic product reviews corpus

6 punkt Word and sentence tokenizer modules

7 treebank Penn Treebank dataset sample

8 twitter\_samples Twitter messages sample

9 universal\_tagset Universal POS tag mapping

10 webtext Web text corpus

11 wordnet WordNet corpus

12 words Word list

1. Tokenization----🡪Word Tokenizer, Sentence Tokenizer

**from nltk.tokenize import word\_tokenize,sent\_tokenize**

example\_text=”Hello,how are you? I am fine. Today sky is pinkish-blue. My mail is @vartak.com”

print(sent\_tokenize(example\_text)) #splits and prints document into sentence

print(word\_tokenize(example\_text)) #splits and prints sentence into words

Even U can use instead of print

for i in word\_tokenize(example\_text):

print (i)

1. Tokenizing punctuations also

from nltk.tokenize import WordPunctTokenizer

example\_text="Hello, how are you? I am fine. Today sky is pinkish-blue. My mail is @vartak.com"

print(WordPunctTokenizer().tokenize(example\_text))

1. Removing stopwords-words which don’t have much meaning/use in search

from nltk.corpus import stopwords

stopw=stopwords.words(‘english’)

print (stopw) #only stopwords from English displayed from

If these stopwords are present in ur input text, then remove them

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

stopw=set(stopwords.words('english'))

example\_text="This is an official site of vartak college."

word=word\_tokenize(example\_text)

filter\_word=[]

for w in word:

if w not in stopw:

filter\_word.append(w)

print (filter\_word)

**OR** u can write

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

stopw=set(stopwords.words('english'))

example\_text="This is an official site of vartak college."

word=word\_tokenize(example\_text)

filter\_word=[w for w in word if not w in stopw] ##in place of following 3 lines

## for w in word:

## if w not in stopw:

## filter\_word.append(w)

print (filter\_word)

1. **Stemming---🡪 finding the stem of word**

Ex. Connect, connected, connecting,connection,connector. – All these variations are impossible to save in dictionary. So only base/stem is saved. We will use Porter Stemmer algorithm to find stem.

## from nltk.tokenize import word\_tokenize

from nltk.stem import PorterStemmer

example\_text=[ "Connect","connected","connecting","connection","connector"]

ps=PorterStemmer()

for w in example\_text:

print(ps.stem(w))

**OR**

from nltk.tokenize import word\_tokenize

from nltk.stem import PorterStemmer

example\_text="It is always needs to be connected with world. otherwise you are connectionless. My network is connecting. I am connector between two friends."

ps=PorterStemmer()

word=word\_tokenize(example\_text)

for w in word:

print(ps.stem(w))