from bs4 import BeautifulSoup

import urllib.request

import nltk

response = urllib.request.urlopen('http://php.net/')

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

tokens = [t for t in text.split()]

freq = nltk.FreqDist(tokens)

for key,val in freq.items():

print (str(key) + ':' + str(val))

Removing Stop words

from bs4 import BeautifulSoup

import urllib.request

import nltk

from nltk.corpus import stopwords

response = urllib.request.urlopen('http://php.net/')

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

tokens = [t for t in text.split()]

clean\_tokens = tokens[:]

sr = stopwords.words('english')

for token in tokens:

if token in stopwords.words('english'):

clean\_tokens.remove(token)

freq = nltk.FreqDist(clean\_tokens)

for key,val in freq.items():

print (str(key) + ':' + str(val))

Tokenize – Sentences as well as words

Sentence Tokenize

from nltk.tokenize import sent\_tokenize

mytext = "Hello Adam, how are you? I hope everything is going well. Today is a good day, see you dude."

print(sent\_tokenize(mytext))

Output

['Hello Adam, how are you?', 'I hope everything is going well.', 'Today is a good day, see you dude.']

Word Tokenize

from nltk.tokenize import word\_tokenize

mytext = "Hello Mr. Adam, how are you? I hope everything is going well. Today is a good day, see you dude."

print(word\_tokenize(mytext))

['Hello', 'Mr.', 'Adam', ',', 'how', 'are', 'you', '?', 'I', 'hope', 'everything', 'is', 'going', 'well', '.', 'Today', 'is', 'a', 'good', 'day', ',', 'see', 'you', 'dude', '.']

Tokenize sentences in French

from nltk.tokenize import sent\_tokenize

mytext = "Bonjour M. Adam, comment allez-vous? J'espère que tout va bien. Aujourd'hui est un bon jour."

print(sent\_tokenize(mytext,"french"))

Find Synonyms

from nltk.corpus import wordnet

syn = wordnet.synsets("pain")

print(syn[0].definition())

print(syn[0].examples())

Lemmas

from nltk.corpus import wordnet

synonyms = []

for syn in wordnet.synsets('Computer'):

for lemma in syn.lemmas():

synonyms.append(lemma.name())

print(synonyms)

Antonyms

from nltk.corpus import wordnet

antonyms = []

for syn in wordnet.synsets("small"):

for l in syn.lemmas():

if l.antonyms():

antonyms.append(l.antonyms()[0].name())

print(antonyms)

STEMMING – Finding the Root word from a series of words

from nltk.stem import PorterStemmer

stemmer = PorterStemmer()

print(stemmer.stem('working'))

Lemmatizing

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print(lemmatizer.lemmatize('increases'))

from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

print(lemmatizer.lemmatize('playing', pos="v"))

print(lemmatizer.lemmatize('playing', pos="n"))

print(lemmatizer.lemmatize('playing', pos="a"))

print(lemmatizer.lemmatize('playing', pos="r"))