1. Introduction

This assignment will help you to consolidate the concepts learnt in the session.

2. Problem Statement

In this assignment students have to find the frequency of words in a webpage. User can

use urllib and BeautifulSoup to extract text from webpage.

Hint:

from bs4 import BeautifulSoup

import urllib.request

import nltk

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

html = response.read()

soup = BeautifulSoup(html,"html5lib")

NOTE: The solution shared through Github should contain

Solution:

# -\*- coding: utf-8 -\*-

"""

Created on Tue Mar 19 21:54:02 2019

@author: ShaniDeva

"""

from bs4 import BeautifulSoup

import urllib.request

import nltk

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

stopwords.words('english')

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

html = response.read()

soup = BeautifulSoup(html,"html5lib")

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

#convert that text into tokens by splitting

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

#tokens = word\_tokenize(text)

clean\_tokens = tokens[:]

sr = stopwords.words('english')

for token in tokens:

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

clean\_tokens.remove(token)

#calculate the frequency of those tokens using Python NLTK.

freq = nltk.FreqDist(tokens)

for key,val in freq.items():

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

#PHP is the highest frequency word- (55 times)





