Session 28: Assignment1

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")

Code:

from bs4 import BeautifulSoup

import urllib.request

import nltk

import re

from collections import Counter

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

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.getText()

words = re.findall(r'\w+', text.lower())

Counter(words).most\_common(10)

Counter(words)

Output:

