In [ ]:

```
#**********************************************************************
#
#              Problem  :  To Find the Frequency of Words in Webpage
#
#**********************************************************************
# Importing the Libraries for NLP
import nltk
from bs4 import BeautifulSoup
%matplotlib inline

# importing urllib package to read the webpage
import urllib.request

# opening the given Url
response = urllib.request.urlopen('http://php.net/')

# getting the web page with the html tags ( unstructured /non formatted )
html = response.read()


# beautifulsoup library is used for reading web pages and gives the page in the format
soup = BeautifulSoup(html,"html5lib")

#removes the space
text = soup.get_text(strip=True)
#print(text)

# getting the words from the text found on the page
tokens = [t for t in text.split()]

# calling the Frequency Distribution function , to get the frequency of words in the web page
freq = nltk.FreqDist(tokens)
maxwd = freq.max()
maxcnt=0

print( " The Words along with their frequency in the web page :\n",'-'*80)

# The frequency is a dictionary with word and the frequency
for key,val in freq.items():
    if (key == maxwd):
        maxcnt = val
```

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

In [ ]:
```
# The words with maximun frequency are
print("\n The word with maximun frequency in the webpage and frequency is :\n",'-'*80)
print(freq.max(),maxcnt)
```

In [ ]:
```
import pandas as pd
print("\n The Top most 10 words used in the webpage are :\n",'-'*80)
# Most common 10 words in the webpage are
words,frq = zip(*freq.most_common(10))
print(words)
```

In [ ]:
```
# Drawing a plot showing the words and their frequency distribution
print(" This Plot shows the Frequency Distribution of top 30 words in webpage \n")
freq.plot(30,cumulative=False)
print ('''\n From The plot , we observe PHP is used maximum on this page , hence the page appears to be
        based on PHP ''')
#*********************************************************************
```