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")
```

In [1]:
```
#Importing the necessary libraries
from bs4 import BeautifulSoup
import urllib.request
import nltk
```

In [2]:
```
response = urllib.request.urlopen('http://php.net/')
html = response.read()
soup = BeautifulSoup(html,"html5lib")
```

In [3]:
```
#Viewing the actual code
soup
```

...

In [4]:
```
#Viewing the text of soup
raw = soup.get_text()
raw
```

...

In [5]:
```
nltk.download('punkt')
words=nltk.word_tokenize(raw)

#removing the singal characters  mostly puncatuations
words=[word for word in words if len(word)>1]

#removing any numbers present in our text
words = [word for word in words if not word.isnumeric()]

#lowercase all words (default stopwords are lowercase too)
words = [word.lower() for word in words]

#calculating frequency distribution
fdist = nltk.FreqDist(words)

#printing the top 30 words with their frequency
print('\t   {}   \t{}'.format("Word","Frequency"))
print("-"*30)
for word , frequency in fdist.most_common(30):
    print('\t',word.center(10,' '),end='')
    print(str(frequency).center(10,' '))
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\idofa\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!

           Word         Frequency
------------------------------
            the       245
            php       152
             of        87
          release      85
            for        81
            this       66
             in        60
             is        56
             to        51
             be        50
            and        49
            can        49
           found       49
           please      44
         downloads     42
          version      42
           source      41
             on        40
            page       33
           8.0.0       29
            list       26
          changes      26
            team       25
           visit       25
            file       20
         candidate     20
         development    18
         availability    18
```

```
test       18
you        18
```

In [ ]: