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.

In [1]:

```
from bs4 import BeautifulSoup
import urllib.request
import nltk
from nltk.corpus import stopwords
```

In [2]:

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

In [3]:

```
soup = BeautifulSoup(html,"html5lib")
```

In [4]:

```
# Get all the text data from webpage
text = soup.get_text(strip=True)
# Split the data into words i.e. tokens
tokens = [t for t in text.split()]
```

In [5]:

```
# Remove stop words from tokens
stw = stopwords.words('english')
clean_tokens = tokens[:]
for token in tokens:
    if token in stopwords.words('english'):
        clean_tokens.remove(token)
```

In [6]:

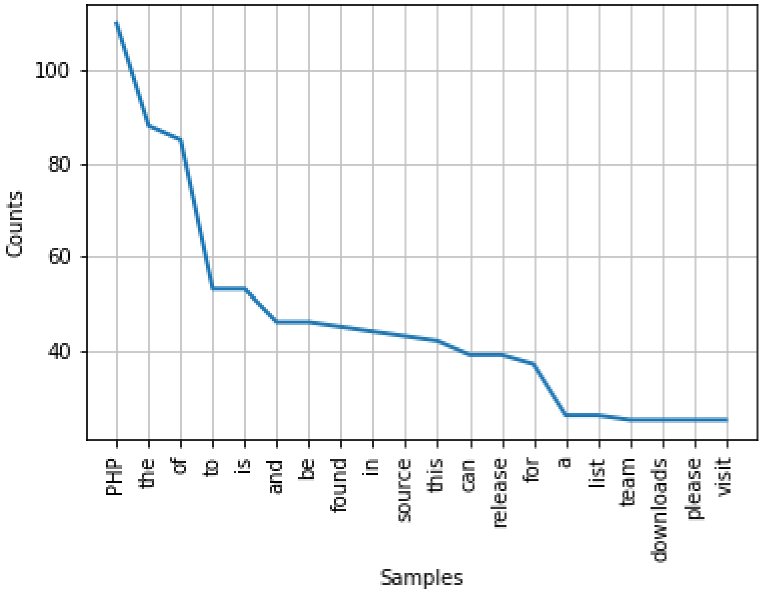
```
# Calculate the frequency of words:
freq = nltk.FreqDist(tokens)
```

In [7]:

```
i = 0
for key,val in freq.items():
    print(str(key)+':'+str(val))
    i = i+1
    if i > 4:
        break
```

```
PHP::1
Hypertext:1
PreprocessorDownloadsDocumentationGet:1
InvolvedHelpGetting:1
StartedIntroductionA:1
```

In [8]:
```
# Plotting top 20 frequently used words
freq.plot(20,cumulative=False)
```



Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x18cf75363c8>
```

In [ ]: