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/ (http://php.net/)')
- html = response.read()
- soup = BeautifulSoup(html,"html5lib")

Importing Required Libraries

In [1]:
```
from bs4 import BeautifulSoup
import urllib.request
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\prashant_gupta1\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[1]:
```
True
```

In [2]:
```
import string
# Opeining the URL Page
response = urllib.request.urlopen('http://php.net/')
# Creating the HTML Page from the URL
html = response.read()
# Using BeautifulSoup to extract text from webpage
soup = BeautifulSoup(html,"html5lib")
#soup = BeautifulSoup(page.content, 'html.parser')

# Generating text from soup and saperating it by space for the new lines
text = soup.get_text(" ", strip=True)

# Removing punctuation from the text, we will not count punctuation here count only for the
exclude = set(string.punctuation)
text = ''.join(ch for ch in text if ch not in exclude)
#print(text)

# Creating the tokens splitting it by space
#print(text)
tokens = [t for t in text.split()]

# Cleaning some of the unwanted words like as, an etc
clean_tokens = tokens[: ]

# Removing the stopwords from the clean tokens by making them in lowercase
sr = stopwords.words('english')
for token in tokens:
    if token.lower() in stopwords.words('english'):
        clean_tokens.remove(token)
# Creating the dictionary with count on clean tokens
freq = nltk.FreqDist(clean_tokens)

# Sorting keys by values in decending order. i.e getting keys of sorted values
freq_sorted_key_by_values = sorted(freq, key=freq.get, reverse=True)

# Printing the data
for key1 in freq_sorted_key_by_values:
    print (str(key1) + ' : ' + str(freq[key1]))

# Here we have not removed the numeric values for now, but those also can be removed
```

```
PHP : 170
release : 75
found : 64
file : 42
version : 33
source : 31
downloads : 31
changes : 31
list : 29
page : 27
Windows : 24
Released : 24
team : 24
please : 24
visit : 24
binaries : 23
read : 22
also : 22
2018 : 21
```