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") NOTE: The solution shared through Github should contain the source code used and the screenshot of the output.

In [1]:
```
import urllib.request
import nltk
from bs4 import BeautifulSoup
response = urllib.request.urlopen('http://php.net/')
html = response.read()
raw = BeautifulSoup(html,"html5lib").get_text()
```

In [3]:
```
nltk.download('punkt')
words = nltk.word_tokenize(raw)

# Removing the single-characters, mostly punctuations
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
for word, frequency in fdist.most_common(30):
    print(u'{}; {}'.format(word, frequency))
```

```
the; 318
php; 186
for; 111
of; 99
in; 98
release; 96
can; 78
and; 77
be; 76
found; 68
is; 66
on; 62
this; 59
file; 46
you; 41
to; 40
please; 40
upgrading; 36
version; 35
downloads; 33
changes; 33
source; 32
or; 32
development; 30
list; 30
page; 28
released; 27
also; 26
windows; 25
team; 25

[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Annonymous-1\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```