Tokenize Text Using Pure Python

First, we will grab some web page content. Then, we will analyze the text to see what the page is about. We will use the [urllib module](https://likegeeks.com/python-programming-basics/" \l "Web-Crawling" \t "_blank) to crawl the web page:

import urllib.request

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

html = response.read()

print (html)

As you can see from the printed output, the result contains a lot of HTML tags that need to be cleaned. We can use BeautifulSoup to clean the grabbed text like this:

from bs4 import BeautifulSoup

import urllib.request

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

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

print (text)

Now, we have clean text from the crawled web page. Awesome, Right?

Finally, let's convert that text into tokens by splitting the text like this:

from bs4 import BeautifulSoup

import urllib.request

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

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

tokens = [t for t in text.split()]

print (tokens)

## Count Word Frequency

The text is much better now. Let's calculate the frequency distribution of those tokens using Python NLTK. There is a function in NLTK called FreqDist() that does the job:

from bs4 import BeautifulSoup

import urllib.request

import nltk

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

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

tokens = [t for t in text.split()]

freq = nltk.FreqDist(tokens)

for key,val in freq.items():

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