assignment_29.1

January 13, 2019

0.1 Assignment 29.1

In this assignment I have used urllib and BeautifulSoup to find the frequency of words in a web-page.

In [19]:
```
## Install bs4
```

In [1]:
```
!pip install bs4
```

```
Collecting bs4
  Downloading https://files.pythonhosted.org/packages/10/ed/7e8b97591f6f456174139ec089c769f89a
Requirement already satisfied: beautifulsoup4 in e:\anaconda\lib\site-packages (from bs4) (4.6
Building wheels for collected packages: bs4
  Running setup.py bdist_wheel for bs4: started
  Running setup.py bdist_wheel for bs4: finished with status 'done'
  Stored in directory: C:\Users\monim\AppData\Local\pip\Cache\wheels\a0\b0\b2\4f80b9456b87abedb
Successfully built bs4
Installing collected packages: bs4
Successfully installed bs4-0.0.1
```

0.2 Import libraries BeautifulSoup,urllib, nltk, download punkt

In [7]:
```
from bs4 import BeautifulSoup
import urllib.request
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\monim\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.
```

Out[7]:
```
True
```

0.3 Get the response from the webpage http://php.net/

```
In [8]: response = urllib.request.urlopen('http://php.net/')
        html = response.read()
        print(html)
```

```
b'<!DOCTYPE html>\n<html xmlns="http://www.w3.org/1999/xhtml" lang="en">\n<head>\n\n  <meta cha
```

0.4 Using BeautifulSoup and ntk tokenized HTML response

```
In [9]: soup = BeautifulSoup(html,"html5lib")
        text_tokens = nltk.tokenize.word_tokenize(soup.get_text())
        print(text_tokens)
```

```
['PHP', ':', 'Hypertext', 'Preprocessor', 'Downloads', 'Documentation', 'Get', 'Involved', 'Hel
```

0.5 Download and import stopwords from nltk

```
In [14]: nltk.download('stopwords')
         from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\monim\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

0.6 Process tokens and if it i snot a stopword, caclulate count and put into a dictionary

```
In [21]: token_count_map ={}
         for token in text_tokens:
             if token not in stopwords.words('english'):
                 if token in token_count_map.keys():
                     count = token_count_map.get(token)
                     count += 1
                     token_count_map[token] = count
                 else:
                     token_count_map[token] = 1
```

0.7 Print each token and its occurrence count

```
In [23]: for token,count in token_count_map.items():
             print (str(token) + ':' + str(count))
```

```
PHP:198
::1
Hypertext:1
```

Preprocessor:1
Downloads:1
Documentation:1
Get:1
Involved:1
Help:1
Getting:1
Started:1
Introduction:2
A:18
simple:1
tutorial:1
Language:2
Reference:2
Basic:2
syntax:1
Types:1
Variables:2
Constants:1
Expressions:1
Operators:1
Control:2
Structures:1
Functions:1
Classes:2
Objects:1
Namespaces:1
Errors:1
Exceptions:2
Generators:1
References:1
Explained:1
Predefined:3
Interfaces:1
Context:1
options:1
parameters:1
Supported:1
Protocols:1
Wrappers:1
Security:4
General:1
considerations:1
Installed:2
CGI:1
binary:1
Apache:1
module:1

```
Session:2
Filesystem:1
Database:3
Error:1
Reporting:1
Using:2
Register:1
Globals:1
User:2
Submitted:1
Data:1
Magic:1
Quotes:1
Hiding:1
Keeping:1
Current:1
Features:1
HTTP:1
authentication:1
Cookies:1
Sessions:1
Dealing:1
XForms:1
Handling:1
file:45
uploads:1
remote:1
files:19
Connection:1
handling:1
Persistent:1
Connections:1
Safe:1
Mode:1
Command:2
line:1
usage:1
Garbage:1
Collection:1
DTrace:1
Dynamic:1
Tracing:1
Function:1
Affecting:1
's:1
Behaviour:1
Audio:1
Formats:1
```

Manipulation:2
Authentication:1
Services:3
Line:1
Specific:2
Extensions:16
Compression:1
Archive:1
Credit:1
Card:1
Processing:3
Cryptography:1
Date:1
Time:1
Related:4
File:1
System:1
Human:1
Character:1
Encoding:1
Support:1
Image:1
Generation:1
Mail:1
Mathematical:1
Non-Text:1
MIME:1
Output:1
Process:1
Other:3
Search:1
Engine:1
Server:1
Text:1
Variable:1
Type:1
Web:1
Windows:25
Only:1
XML:1
GUI:1
Keyboard:1
Shortcuts:1
?:1
This:13
help:1
j:1
Next:2

menu:2
item:2
k:1
Previous:2
g:6
p:1
man:2
page:28
n:1
G:1
Scroll:2
bottom:1
top:1
h:1
Goto:2
homepage:1
search:2
(:2
current:1
):2
/:1
Focus:1
box:1
popular:2
general-purpose:1
scripting:1
language:1
especially:1
suited:1
web:1
development:13
.:227
Fast:1
,:92
flexible:1
pragmatic:1
powers:1
everything:1
blog:1
websites:1
world:1
Download:1
5.6.40ůRelease:1
NotesůUpgrading:4
7.1.26ůRelease:1
7.2.14ůRelease:1
7.3.1ůRelease:1
10:4

Jan:4
2019:5
5.6.40:4
Released:24
The:75
team:25
announces:12
immediate:12
availability:12
security:8
release:84
Several:1
bugs:6
fixed:2
All:12
5.6:4
users:12
encouraged:12
upgrade:9
version:36
For:43
source:32
downloads:32
please:25
visit:25
binaries:24
found:68
windows.php.net/download/:7
list:30
changes:33
recorded:7
ChangeLog:7
Please:16
note:2
according:2
support:2
timelines:2
last:3
scheduled:2
branch:4
There:2
may:4
additional:2
discover:2
important:2
issues:16
warrant:2
otherwise:2

final:3
one:2
If:2
installation:2
based:2
good:2
time:2
start:2
making:2
plans:2
7.1:4
7.2:4
7.3:16
7.1.26:3
Release:15
Announcement:1
also:26
contains:4
several:4
bug:22
fixes:5
7.3.1:3
7.2.14:3
minor:1
06:3
Dec:2
2018:16
7.0.33:4
Five:1
security-related:1
7.0:4
7.1.25:3
22:1
Nov:2
7.3.0RC6:3
glad:13
announce:13
presumably:1
7.3.0:23
pre-release:6
rough:13
outline:13
cycle:13
specified:13
Wiki:13
download:18
sources:17
windows.php.net/qa/:17

carefully:18
test:19
report:18
reporting:13
system:18
THIS:17
IS:17
DEVELOPMENT:17
PREVIEW:17
-:19
DO:18
NOT:18
USE:17
IT:17
IN:17
PRODUCTION:17
!:18
information:18
new:18
features:18
read:23
NEWS:18
UPGRADING:18
complete:18
upgrading:18
notes:18
Internal:8
listed:8
UPGRADING.INTERNALS:8
These:18
archive:18
next:20
would:13
GA:1
planned:18
December:1
6th:1
signatures:13
manifest:13
QA:13
site:13
Thank:18
helping:18
us:18
make:18
better:18
08:1
7.3.0RC5:3

RC6:1
November:2
22nd:1
25:1
Oct:3
7.3.0RC4:3
RC5:1
8th:1
11:1
7.3.0RC3:3
RC4:2
October:4
25th:1
28:2
Sep:3
7.3.0RC2:3
RC3:2
11th:1
13:1
7.3.0RC1:3
RC2:1
September:3
27th:1
30:1
Aug:5
7.3.0.beta3:1
seventh:1
7.3.0beta3:2
RC1:1
13th:1
16:1
7.3.0.beta2:1
sixth:1
7.3.0beta2:2
Beta:7
3:13
August:4
30th:1
02:1
7.3.0.beta1:1
fifth:1
7.3.0beta1:2
2:6
16th:1
19:1
Jul:3
7.3.0alpha4:3
fourth:2

1:8
2nd:1
05:1
alpha:3
third:3
Alpha:12
July:3
19th:1
21:2
Jun:2
second:2
5:1
07:1
first:4
starts:1
use:1
production:1
early:1
June:1
01:1
Feb:1
7.2.2:3
bugfix:1
included:1
12:1
2017:5
7.2.0:19
Candidate:14
4:2
incompatibilities:5
tracking:5
announced:2
26th:1
You:5
full:5
releases:5
wiki:5
12th:1
31:1
released:3
14th:1
17:1
beta:2
31th:1
improvements:1
relative:1
20th:1
Older:1

News:1
Entries:1
Upcoming:1
conferencesSunshinePHP:1
2019Dutch:2
Conference:4
2019International:1
Spring:1
Edition:1
Conferences:1
calling:1
papersPHPKonf:1
Istanbul:1
CfP:1
open:1
Group:2
Events:1
Special:1
Thanks:1
Social:1
media:1
@:1
official_php:1
Copyright:1
I':1
2001-2019:1
My:1
PHP.net:2
Contact:1
sites:2
Mirror:1
Privacy:1
policy:1