Data Science Masters :Assignment 28

Problem:

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.

Solution:

Importing Libraries...

In [1]:

```
from bs4 import BeautifulSoup
import urllib.request
import nltk
nltk.download('stopwords')
nltk.download('wordnet')
from nltk.corpus import stopwords
from nltk.corpus import stopwords
from textblob import TextBlob
from textblob import Word
from nltk.stem import PorterStemmer
import pandas as pd
import matplotlib.pyplot as plt

from subprocess import check_output
from wordcloud import WordCloud, STOPWORDS
```

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

In [2]:

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

In [3]:

```
print(html)
```

```
b'<!DOCTYPE html>\n<html xmlns="http://www.w3.org/1999/xhtml" lang="en">\n
<head>\n\n  <meta charset="utf-8">\n  <meta name="viewport" content="width
=device-width, initial-scale=1.0">\n\n  <title>PHP: Hypertext Preprocessor
</title>\n\n <link rel="shortcut icon" href="http://php.net/favicon.ico">
\n <link rel="search" type="application/opensearchdescription+xml" href="h
ttp://php.net/phpnetimprovedsearch.src" title="Add PHP.net search">\n <lin
k rel="alternate" type="application/atom+xml" href="http://php.net/release
s/feed.php" title="PHP Release feed">\n <link rel="alternate" type="applic
ation/atom+xml" href="http://php.net/feed.atom" title="PHP: Hypertext Prep
rocessor">\n\n <link rel="canonical" href="http://php.net/index.php">\n <l
ink rel="shorturl" href="http://php.net/index">\n <link rel="alternate" hr
ef="http://php.net/index" hreflang="x-default">\n\n\n\n<link rel="styleshe
et" type="text/css" href="/cached.php?t=1539771603&amp;f=/fonts/Fira/fira.
css" media="screen">\n<link rel="stylesheet" type="text/css" href="/cache
d.php?t=1539765004&amp;f=/fonts/Font-Awesome/css/fontello.css" media="scre
en">\n<link rel="stylesheet" type="text/css" href="/cached.php?t=154042560
3&amp;f=/styles/theme-base.css" media="screen">\n<link rel="stylesheet" ty
pe="text/css" href="/cached.php?t=1540425603&amp;f=/styles/theme-medium.cs
s" media="screen">\n<link rel="stylesheet" type="text/css" href="/cached.p
```

In [4]:

```
#use BeautifulSoup to clean the grabbed text data which is at HTML
soup = BeautifulSoup(html,"html5lib")
text = soup.get_text(strip=True)
print(text)
```

```
HP
     5.6.39. This is a security release. Several security bugs have been f
ixed
     in this release.

     All PHP 5.6 users are encouraged to upgrade to this version.For sourc
e downloads of PHP 5.6.39 please visit ourdownloads page,
     Windows source and binaries can be found onwindows.php.net/download/.
     The list of changes is recorded in theChangeLog.Please note that acco
rding to thePHP version
support timelines,
        PHP 5.6.39 is the last scheduled release of PHP 5.6 branch. There
may be additional release if we
discover
        important security issues that warrant it, otherwise this release
will be the final one in the PHP
5.6 branch.
        If your PHP installation is based on PHP 5.6, it may be a good tim
e to start making the plans for
the upgrade
```

In [5]:

```
#convert that text into tokens by splitting the text
tokens = [t for t in text.split()]
print(tokens)
```

```
inaries', 'can', 'be', 'found', 'onwindows.php.net/download/.', 'The', 'li
st', 'of', 'changes', 'is', 'recorded', 'in', 'theChangeLog.06', 'Dec', '2
018PHP', '7.2.13', 'ReleasedThe', 'PHP', 'development', 'team', 'announce
s', 'the', 'immediate', 'availability', 'of', 'PHP', '7.2.13.', 'This', 'i
s', 'a', 'security', 'release.All', 'PHP', '7.2', 'users', 'are', 'encoura
ged', 'to', 'upgrade', 'to', 'this', 'version.For', 'source', 'downloads',
'of', 'PHP', '7.2.13', 'please', 'visit', 'ourdownloads', 'page,', 'Window
s', 'source', 'and', 'binaries', 'can', 'be', 'found', 'onwindows.php.net/
download/.', 'The', 'list', 'of', 'changes', 'is', 'recorded', 'in', 'theC
hangeLog.06', 'Dec', '2018PHP', '5.6.39', 'ReleasedThe', 'PHP', 'developme
nt', 'team', 'announces', 'the', 'immediate', 'availability', 'of', 'PHP',
'5.6.39.', 'This', 'is', 'a', 'security', 'release.', 'Several', 'securit
y', 'bugs', 'have', 'been', 'fixed', 'in', 'this', 'release.', 'All', 'PH
P', '5.6', 'users', 'are', 'encouraged', 'to', 'upgrade', 'to', 'this', 'v
ersion.For', 'source', 'downloads', 'of', 'PHP', '5.6.39', 'please', 'visi
t', 'ourdownloads', 'page,', 'Windows', 'source', 'and', 'binaries', 'ca
n', 'be', 'found', 'onwindows.php.net/download/.', 'The', 'list', 'of', 'c
hanges', 'is', 'recorded', 'in', 'theChangeLog.Please', 'note', 'that', 'a
ccording', 'to', 'thePHP', 'version', 'support', 'timelines,', 'PHP', '5.
6.39', 'is', 'the', 'last', 'scheduled', 'release', 'of', 'PHP', '5.6', 'b
```

Frequency Distribution

In [6]:

```
#calculate the frequency distribution using Python NLTK
freq = nltk.FreqDist(tokens)
#Loop through and print
for key,val in freq.items():
    print(str(key) + ':' + str(val))
```

```
PHP::1
Hypertext:1
PreprocessorDownloadsDocumentationGet:1
InvolvedHelpGetting:1
StartedIntroductionA:1
simple:1
tutorialLanguage:1
ReferenceBasic:1
syntaxTypesVariablesConstantsExpressionsOperatorsControl:1
StructuresFunctionsClasses:1
and:80
ObjectsNamespacesErrorsExceptionsGeneratorsReferences:1
ExplainedPredefined:1
VariablesPredefined:1
ExceptionsPredefined:1
Interfaces:1
ClassesContext:1
options:1
parametersSupported:1
Protocols:1
WrappersSecurityIntroductionGeneral:1
considerationsInstalled:1
as:2
CGI:1
binaryInstalled:1
an:2
Apache:1
moduleSession:1
SecurityFilesystem:1
SecurityDatabase:1
SecurityError:1
ReportingUsing:1
Register:1
GlobalsUser:1
Submitted:1
DataMagic:1
QuotesHiding:1
PHPKeeping:1
CurrentFeaturesHTTP:1
authentication:1
with:4
PHPCookiesSessionsDealing:1
XFormsHandling:1
file:1
uploadsUsing:1
remote:1
filesConnection:1
handlingPersistent:1
Database:1
ConnectionsSafe:1
ModeCommand:1
line:1
usageGarbage:1
```

CollectionDTrace:1
Dynamic:1
TracingFunction:1
ReferenceAffecting:1
PHP's:1
BehaviourAudio:1
Formats:1
ManipulationAuthentication:1
ServicesCommand:1
Line:1
Specific:2
ExtensionsCompression:1
Archive:1
ExtensionsCredit:1
Card:1
ProcessingCryptography:1
ExtensionsDatabase:1
ExtensionsDate:1
Time:1
Related:4
ExtensionsFile:1
System:1
ExtensionsHuman:1
Language:1
Character:1
Encoding:1
SupportImage:1
Processing:1
GenerationMail:1
ExtensionsMathematical:1
ExtensionsNon-Text:1
MIME:1
OutputProcess:1
Control:1
ExtensionsOther:2
Basic:1
ServicesSearch:1
Engine:1
ExtensionsServer:1
ExtensionsSession:1
ExtensionsText:1
ProcessingVariable:1
Type:1
ExtensionsWeb:1
ServicesWindows:1
Only:1
ExtensionsXML:1
ManipulationGUI:1
ExtensionsKeyboard:1
Shortcuts?This:1
helpjNext:1
menu:2
itemkPrevious:1
itemg:1
pPrevious:1
man:2
pageg:1
nNext:1
pageGScroll:1
to:43
bottomg:1

gScroll:1
topg:1
hGoto:1
homepageg:1
sGoto:1
search(current:1
page)/Focus:1
search:1
boxPHP:1
is:50
a:26
popular:2
general-purpose:1
scripting:1
language:1
that:5
especially:1
suited:1
web:1
development.Fast,:1
flexible:1
pragmatic,:1
PHP:153
powers:1
everything:1
from:1
your:3
blog:1
the:126
most:1
websites:1
in:71
world.Download5.6.39·Release:1
Notes·Upgrading7.0.33·Release:1
Notes·Upgrading7.1.25·Release:1
Notes·Upgrading7.2.13·Release:1
Notes·Upgrading7.3.0·Release:1
Notes·Upgrading06:1
Dec:5
2018PHP:20
7.0.33:3
ReleasedThe:23
development:12
team:25
announces:12
immediate:12
availability:12
of:101
7.0.33.:1
Five:1
security-related:1
issues:16
were:1
fixed:2
this:29
release.:3
All:7
7.0:3
users:11
are:19
encouraged:11

upgrade:8
version.For:7
source:31
downloads:25
please:25
visit:25
ourdownloads:6
page,:6
Windows:24
binaries:24
can:78
be:79
found:68
onwindows.php.net/download/.:6
The:20
list:31
changes:15
recorded:7
theChangeLog.Please:2
note:2
according:2
thePHP:16
version:21
support:2
timelines,:2
last:3
scheduled:2
release:77
branch.:4
There:2
may:4
additional:2
if:2
we:2
discover:2
important:2
security:6
warrant:2
it,:2
otherwise:2
will:7
final:3
one:2
If:2
installation:2
based:2
on:27
7.0,:1
it:4
good:2
time:2
start:2
making:2
plans:2
for:71
7.1,:2
7.2:4
or:20
7.3.06:2
7.1.25:2
7.1.25.:1

This:12
release.All:3
7.1:2
theChangeLog.06:2
7.2.13:2
7.2.13.:1
5.6.39:3
5.6.39.:1
Several:1
bugs:6
have:1
been:1
5.6:3
5.6,:1
7.3.0:26
7.3.0.:1
marks:1
third:4
feature:1
update:1
7:1
series.PHP:1
comes:1
numerous:1
improvements:2
new:20
features:20
such:1
asFlexible:1
Heredoc:1
Nowdoc:1
SyntaxPCRE2:1
MigrationMultiple:1
MBString:1
ImprovementsLDAP:1
Controls:1
SupportImproved:1
FPM:1
LoggingWindows:1
File:1
Deletion:1
ImprovementsSeveral:1
DeprecationsFor:1
ourdownloadspage:1
Windowssite.:1
theChangeLog.Themigration:1
guideis:1
available:1
Manual.:1
Please:1
consult:1
detailed:1
backward:1
incompatible:1
changes.Many:1
thanks:1
all:1
contributors:1
supporters!22:1
Nov:3
7.3.0RC6:2

glad:13
announce:13
presumably:1
pre-release,:6
7.3.0RC6.:1
rough:13
outline:13
7.3:13
cycle:12
specified:13
Wiki.For:13
thedownload:13
page.:12
sources:17
onwindows.php.net/qa/.Please:12
carefully:13
test:19
report:18
any:18
thebug:18
reporting:13
system.THIS:17
IS:17
A:17
DEVELOPMENT:17
PREVIEW:17
-:19
DO:18
NOT:18
USE:17
IT:17
IN:17
PRODUCTION!For:17
more:17
information:18
other:18
changes,:18
you:36
read:23
theNEWSfile,:18
theUPGRADINGfile:18
complete:18
upgrading:18
notes.:18
Internal:8
listed:8
theUPGRADING.INTERNALSfile.:8
These:18
files:18
also:23
archive.The:13
next:20
would:13
(GA),:1
planned:18
December:1
6th.The:1
signatures:13
inthe:13
manifestor:13
onthe:13

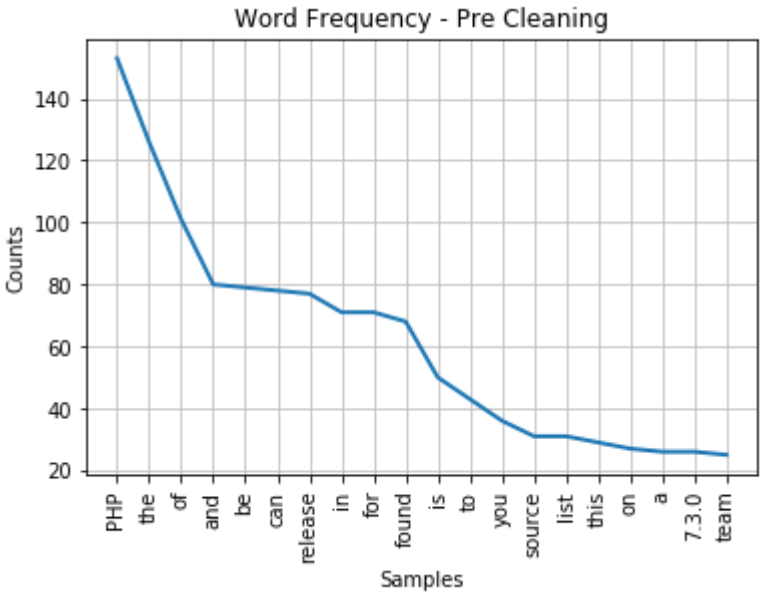
QA:13
site.Thank:13
helping:18
us:18
make:18
better.08:1
7.1.24:3
ReleasedPHP:2
Release:15
AnnouncementThe:1
7.1.24.:1
bugfix:2
theChangeLog.08:1
7.3.0RC5:2
7.3.0RC5.:1
RC6,:1
November:2
22nd.The:1
better.25:1
Oct:3
7.3.0RC4:2
7.3.0RC4.:1
RC5,:1
8th.The:1
better.11:1
7.3.0RC3:2
7.3.0RC3.:1
RC4,:1
October:2
25th.The:1
better.28:2
Sep:3
7.3.0RC2:2
7.3.0RC2.:1
RC3,:1
11th.The:1
better.13:1
7.3.0RC1:2
7.3.0RC1.:1
RC2,:1
September:2
27th.The:1
better.30:1
Aug:5
7.3.0.beta3:1
seventh:1
version,:7
7.3.0beta3.:1
7.3.0beta3:1
RC1,:1
13th.The:1
better.16:1
7.3.0.beta2:1
sixth:1
7.3.0beta2.:1
7.3.0beta2:1
Beta:7
3,:2
August:3
30th.The:1
better.02:1

7.3.0.beta1:1
fifth:1
7.3.0beta1.:1
7.3.0beta1:1
2,:2
16th.The:1
better.19:1
Jul:3
7.3.0alpha4:2
fourth:2
7.3.0alpha4.:1
1,:2
2nd.The:1
better.05:1
alpha:3
3:8
Alpha:12
3.:3
July:2
19th.The:1
better.21:1
Jun:2
2:2
second:2
2.:2
5.The:1
better.07:1
1:4
first:4
1.:2
starts:1
cycle,:1
which:1
page.Please:1
system.Please:1
use:1
production,:1
early:1
June:1
21.The:1
better.01:1
Feb:1
7.2.2:2
7.2.2.:1
release,:1
several:1
bug:1
fixes:2
included.All:1
theChangeLog.12:1
2017PHP:5
7.2.0:15
Candidate:14
4:2
RC4.:1
7.2.0.:4
carefully,:5
incompatibilities:5
tracking:5
archive.For:5
thedownloadpage,:5

atwindows.php.net/qa/.The:4
announced:2
26th:1
October.:2
You:5
full:5
releases:5
onour:4
wiki.Thank:4
RC3.:1
12th:1
better.31:1
released:3
14th:1
September.:1
better.17:1
beta:2
31th:1
August.:1
better.06:1
contains:1
relative:1
onwindows.php.net/qa/.The:1
20th:1
July.:1
ourwiki.Thank:1
better.Older:1
News:1
EntriesUpcoming:1
conferencesSunshinePHP:1
2019Dutch:2
Conference:4
2019International:1
2019:1
Spring:1
EditionConferences:1
calling:1
papersPHPKonf:1
Istanbul:1
CfP:1
open!User:1
Group:1
EventsSpecial:1
ThanksSocial:1
media@official_phpCopyright:1
©:1
2001-2018:1
GroupMy:1
PHP.netContactOther:1
PHP.net:1
sitesMirror:1
sitesPrivacy:1
policy:1

In [7]:

```
#plot a graph
freq.plot(20, cumulative=False,title='Word Frequency - Pre Cleaning')
```



Analyzing Data

In [8]:

```
# Creating Dataframe..
word_df = pd.DataFrame()
word_df['Word']=freq.keys()
word_df['Count']=freq.values()
```

In [9]:

```
# Printing first five rows
word_df.head()
```

Out[9]:

	Word	Count
0	PHP:	1
1	Hypertext	1
2	PreprocessorDownloadsDocumentationGet	1
3	InvolvedHelpGetting	1
4	StartedIntroductionA	1

In [10]:

```
# Calculating characters length
word_df['charLength'] = word_df['Word'].str.len()
word_df.head(5)
```

Out[10]:

	Word	Count	charLength
0	PHP:	1	4
1	Hypertext	1	9
2	PreprocessorDownloadsDocumentationGet	1	37
3	InvolvedHelpGetting	1	19
4	StartedIntroductionA	1	20

In [11]:

```
# finding stopwords
stop = stopwords.words('english')
word_df['stopwords'] = word_df['Word'].apply(lambda x: len([x for x in x.split() if x in st
word_df.loc[(word_df.stopwords.values!=0)].head()
```

Out[11]:

	Word	Count	charLength	stopwords
10	and	80	3	1
22	as	2	2	1
25	an	2	2	1
40	with	4	4	1
112	to	43	2	1

In [12]:

```
# finding Numbers
word_df['Numbers'] = word_df['Word'].apply(lambda x: len([x for x in x.split() if x.isdigit
word_df.loc[(word_df.Numbers.values!=0)].head()
```

Out[12]:

	Word	Count	charLength	stopwords	Numbers
256	7	1	1	0	1
434	3	8	1	0	1
441	2	2	1	0	1
446	1	4	1	0	1
472	4	2	1	0	1

In [13]:

```
# finding special characters
word_df['speicalChar'] = word_df['Word'].apply(lambda x: len([x for x in x.split() if x.enc
word_df.loc[(word_df.speicalChar.values!=0)].head()
```

Out[13]:

	Word	Count	charLength	stopwords	Numbers	speicalChar
0	PHP:	1	4	0	0	1

Data Pre-Processing

In [14]:

```
# removing punctuation marks
word_df['Word'] = word_df['Word'].str.replace('[^\w\s]','')
word_df['speicalChar'] = word_df['Word'].apply(lambda x: len([x for x in x.split() if x.enc
word_df.head(1)
```

Out[14]:

	Word	Count	charLength	stopwords	Numbers	speicalChar
0	PHP	1	4	0	0	0

In [15]:

```
# converting to lower cases
word_df['Word'] = word_df['Word'].apply(lambda x: " ".join(x.lower() for x in x.split()))
word_df.head()
```

Out[15]:

	Word	Count	charLength	stopwords	Numbers	speicalChar
0	php	1	4	0	0	0
1	hypertext	1	9	0	0	0
2	preprocessordownloadsdocumentationget	1	37	0	0	0
3	involvedhelpgetting	1	19	0	0	0
4	startedintroductiona	1	20	0	0	0

In [16]:

```
# removing stopwords
word_df['Word'] = word_df['Word'].apply(lambda x: " ".join(x for x in x.split() if x not in
word_df.shape
```

Out[16]:

```
(533, 6)
```

In [17]:

```
# Lemmatization -  word into its root word
word_df['Word'] = word_df['Word'].apply(lambda x: " ".join([Word(word).lemmatize() for word
```

In [18]:

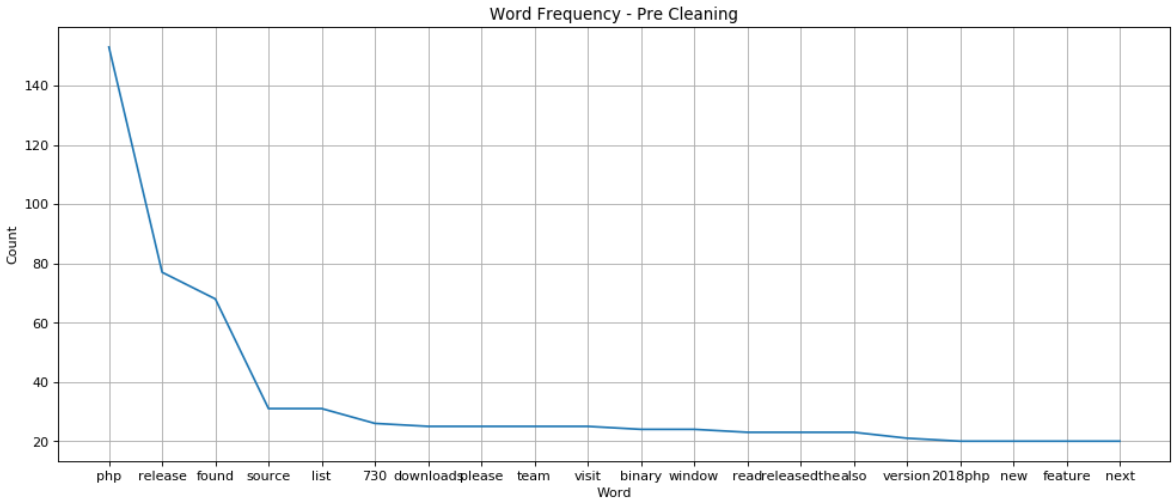
```
#Filtering top words
word_df_top=(word_df.loc[(word_df.Word.values!='')].sort_values(by=['Count'],ascending=Fals
word_df_top
```

Out[18]:

	Word	Count	charLength	stopwords	Numbers	speicalChar
136	php	153	3	0	0	0
202	release	77	7	0	0	0
187	found	68	5	0	0	0
177	source	31	6	0	0	0
190	list	31	4	0	0	0
250	730	26	5	0	0	0
178	downloads	25	9	0	0	0
179	please	25	6	0	0	0
157	team	25	4	0	0	0
180	visit	25	5	0	0	0
184	binary	24	8	0	0	0
183	window	24	7	0	0	0
335	read	23	4	0	0	0
155	releasedthe	23	11	0	0	0
346	also	23	4	0	0	0
197	version	21	7	0	0	0
153	2018php	20	7	0	0	0
261	new	20	3	0	0	0
262	feature	20	8	0	0	0
348	next	20	4	0	0	0

In [19]:

```
# plot a graph - cleaned data
from matplotlib.pyplot import figure
figure(num=None, figsize=(15, 6), dpi=80, facecolor='w', edgecolor='k')
plt.plot(word_df_top.Word.values, word_df_top.Count.values)
plt.xlabel('Word')
plt.ylabel('Count')
plt.title('Word Frequency - Pre Cleaning')
plt.grid()
plt.show()
```



In [22]:

```
# Word Cloud Visulaization
from matplotlib.pyplot import figure
figure(num=None, figsize=(10,5), dpi=80, facecolor='w', edgecolor='k')
stopwords = set(STOPWORDS)
wordcloud = WordCloud(
                          background_color='white',
                          stopwords=stopwords,
                          max_words=40,
                          max_font_size=50,
                          random_state=0
                         ).generate(str(word_df['Word']))
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
```