In [1]:

```
from bs4 import BeautifulSoup
import urllib.request
import nltk
nltk.download('stopwords')
```

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

Out[1]:

```
True
```

In [*]:

```
We will use beautiful soup to crawl our webpage text of HTML tags.
```

In [2]:

```
response = urllib.request.urlopen('http://php.net/')
html = response.read()
soup = BeautifulSoup(html,"html5lib")
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="https://www.php.net/favicon.i
co">\n <link rel="search" type="application/opensearchdescription+xml" hre
f="http://php.net/phpnetimprovedsearch.src" title="Add PHP.net search">\n
<link rel="alternate" type="application/atom+xml" href="https://www.php.ne
t/releases/feed.php" title="PHP Release feed">\n <link rel="alternate" typ
e="application/atom+xml" href="https://www.php.net/feed.atom" title="PHP:
Hypertext Preprocessor">\n\n <link rel="canonical" href="https://www.php.n
et/index.php">\n <link rel="shorturl" href="https://www.php.net/index">\n
<link rel="alternate" href="https://www.php.net/index" hreflang="x-defaul
t">\n\n\n\n<link rel="stylesheet" type="text/css" href="/cached.php?t=1539
771603&amp;f=/fonts/Fira/fira.css" media="screen">\n<link rel="stylesheet"
type="text/css" href="/cached.php?t=1539765004&amp;f=/fonts/Font-Awesome/c
ss/fontello.css" media="screen">\n<link rel="stylesheet" type="text/css" h
ref="/cached.php?t=1540425603&amp;f=/styles/theme-base.css" media="scree
n">\n<link rel="stylesheet" type="text/css" href="/cached.php?t=1540425603
&amp;f=/styles/theme-medium.css" media="screen">\n<link rel="stylesheet" t
ype="text/css" href="/cached.php?t=1429259403&amp;f=/styles/home.css" medi
```

In [3]:

```
text = soup.get_text(strip = True)
print(text)
```

```
PHP: Hypertext PreprocessorDownloadsDocumentationGet InvolvedHelpGetting S
tartedIntroductionA simple tutorialLanguage ReferenceBasic syntaxTypesVari
ablesConstantsExpressionsOperatorsControl StructuresFunctionsClasses and O
bjectsNamespacesErrorsExceptionsGeneratorsReferences ExplainedPredefined V
ariablesPredefined ExceptionsPredefined Interfaces and ClassesContext opti
ons and parametersSupported Protocols and WrappersSecurityIntroductionGene
ral considerationsInstalled as CGI binaryInstalled as an Apache moduleSess
ion SecurityFilesystem SecurityDatabase SecurityError ReportingUsing Regis
ter GlobalsUser Submitted DataMagic QuotesHiding PHPKeeping CurrentFeature
sHTTP authentication with PHPCookiesSessionsDealing with XFormsHandling fi
le uploadsUsing remote filesConnection handlingPersistent Database Connect
ionsSafe ModeCommand line usageGarbage CollectionDTrace Dynamic TracingFun
ction ReferenceAffecting PHP's BehaviourAudio Formats ManipulationAuthenti
cation ServicesCommand Line Specific ExtensionsCompression and Archive Ext
ensionsCredit Card ProcessingCryptography ExtensionsDatabase ExtensionsDat
e and Time Related ExtensionsFile System Related ExtensionsHuman Language
and Character Encoding SupportImage Processing and GenerationMail Related
ExtensionsMathematical ExtensionsNon-Text MIME OutputProcess Control Exten
sionsOther Basic ExtensionsOther ServicesSearch Engine ExtensionsServer Sp
```

In [4]:

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

```
['PHP:', 'Hypertext', 'PreprocessorDownloadsDocumentationGet', 'InvolvedHe
lpGetting', 'StartedIntroductionA', 'simple', 'tutorialLanguage', 'Referen
ceBasic', 'syntaxTypesVariablesConstantsExpressionsOperatorsControl', 'Str
ucturesFunctionsClasses', 'and', 'ObjectsNamespacesErrorsExceptionsGenerat
orsReferences', 'ExplainedPredefined', 'VariablesPredefined', 'ExceptionsP
redefined', 'Interfaces', 'and', 'ClassesContext', 'options', 'and', 'para
metersSupported', 'Protocols', 'and', 'WrappersSecurityIntroductionGenera
l', 'considerationsInstalled', 'as', 'CGI', 'binaryInstalled', 'as', 'an',
'Apache', 'moduleSession', 'SecurityFilesystem', 'SecurityDatabase', 'Secu
rityError', 'ReportingUsing', 'Register', 'GlobalsUser', 'Submitted', 'Dat
aMagic', 'QuotesHiding', 'PHPKeeping', 'CurrentFeaturesHTTP', 'authenticat
ion', 'with', 'PHPCookiesSessionsDealing', 'with', 'XFormsHandling', 'fil
e', 'uploadsUsing', 'remote', 'filesConnection', 'handlingPersistent', 'Da
tabase', 'ConnectionsSafe', 'ModeCommand', 'line', 'usageGarbage', 'Collec
tionDTrace', 'Dynamic', 'TracingFunction', 'ReferenceAffecting', "PHP's",
'BehaviourAudio', 'Formats', 'ManipulationAuthentication', 'ServicesComman
d', 'Line', 'Specific', 'ExtensionsCompression', 'and', 'Archive', 'Extens
ionsCredit', 'Card', 'ProcessingCryptography', 'ExtensionsDatabase', 'Exte
nsionsDate', 'and', 'Time', 'Related', 'ExtensionsFile', 'System', 'Relate
```

Count word Frequency

In [6]:

```
from nltk.corpus import stopwords
sr= stopwords.words('english')
clean_tokens = tokens[:]
for token in tokens:
    if token in stopwords.words('english'):

        clean_tokens.remove(token)
freq = nltk.FreqDist(clean_tokens)
for key,val in freq.items():
    print(str(key) + ':' + str(val))
freq.plot(20, cumulative=False)
```

```
PHP::1
Hypertext:1
PreprocessorDownloadsDocumentationGet:1
InvolvedHelpGetting:1
StartedIntroductionA:1
simple:1
tutorialLanguage:1
ReferenceBasic:1
syntaxTypesVariablesConstantsExpressionsOperatorsControl:1
StructuresFunctionsClasses:1
ObjectsNamespacesErrorsExceptionsGeneratorsReferences:1
ExplainedPredefined:1
VariablesPredefined:1
ExceptionsPredefined:1
Interfaces:1
ClassesContext:1
options:1
parametersSupported:1
Protocols:1
WrappersSecurityIntroductionGeneral:1
considerationsInstalled:1
CGI:1
binaryInstalled:1
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
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
```

```
bottomg:1
gScroll:1
topg:1
hGoto:1
homepageg:1
sGoto:1
search(current:1
page)/Focus:1
search:1
boxPHP:1
popular:2
general-purpose:1
scripting:1
language:1
especially:1
suited:1
web:1
development.Fast,:1
flexible:1
pragmatic,:1
PHP:133
powers:1
everything:1
blog:1
websites:1
world.Download7.1.28·Release:1
Notes·Upgrading7.2.17·Release:1
Notes·Upgrading7.3.4·Release:1
Notes·Upgrading04:1
Apr:3
2019PHP:6
7.1.28:2
ReleasedThe:22
development:12
team:25
announces:12
immediate:12
availability:12
7.1.28.:1
This:13
security:6
release.All:1
7.1:2
users:12
encouraged:12
upgrade:7
version.For:8
source:32
downloads:25
please:25
visit:25
ourdownloads:7
page,:7
Windows:24
binaries:24
found:68
onwindows.php.net/download/.:7
The:20
list:30
changes:15
recorded:7
```

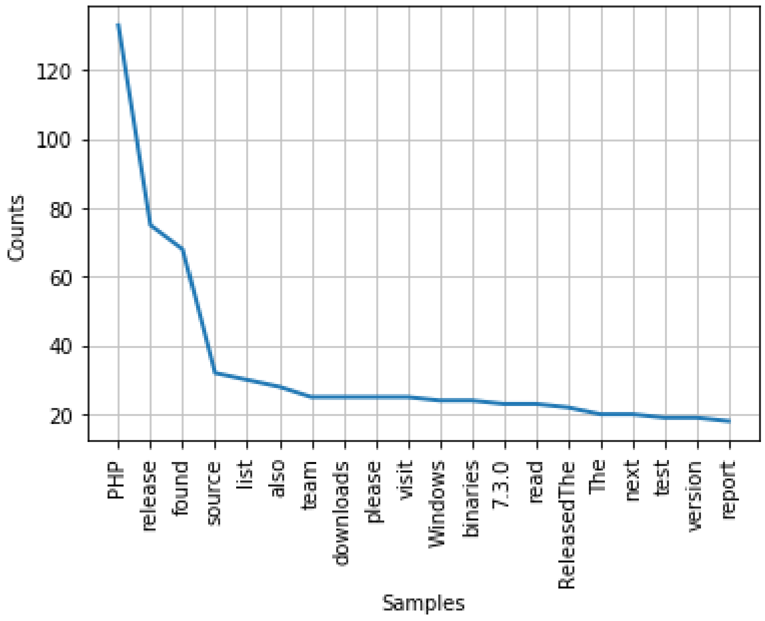
```
theChangeLog.04:2
7.2.17:2
Release:16
AnnouncementThe:2
7.2.17.:1
release:75
also:28
contains:6
several:6
minor:2
bug:6
fixes.All:5
7.2:3
7.3.4:2
7.3.4.:1
7.3:15
theChangeLog.07:3
Mar:3
7.1.27:2
7.1.27.:1
7.2.16:2
7.2.16.:1
7.3.3:2
7.3.3.:1
theChangeLog.22:1
Nov:2
2018PHP:14
7.3.0RC6:2
glad:13
announce:13
presumably:1
last:1
7.3.0:23
pre-release,:6
7.3.0RC6.:1
rough:13
outline:13
cycle:12
specified:13
thePHP:13
Wiki.For:13
thedownload:13
page.:12
sources:17
onwindows.php.net/qa/.Please:12
carefully:13
test:19
version:19
report:18
issues:13
thebug:18
reporting:13
system.THIS:17
IS:17
A:17
DEVELOPMENT:17
PREVIEW:17
-:17
DO:18
NOT:18
USE:17
```

```
IT:17
IN:17
PRODUCTION!For:17
information:18
new:18
features:18
changes,:18
read:23
theNEWSfile,:18
theUPGRADINGfile:18
complete:18
upgrading:18
notes.:18
Internal:8
listed:8
theUPGRADING.INTERNALSfile.:8
These:18
files:18
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
7.3.0RC5:2
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
third:3
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
ReleasedPHP:1
first:4
1.:2
starts:1
cycle,: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
bugfix:1
release,: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
All:5
carefully,:5
bugs: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
final:1
beta:2
31th:1
August.:1
better.06:1
improvements:1
relative:1
onwindows.php.net/qa/.The:1
20th:1
July.:1
ourwiki.Thank:1
better.Older:1
News:1
EntriesUpcoming:1
conferencesPHPConf.Asia:1
2019SymfonyCon:1
Amsterdam:1
2019SymfonyLive:2
Berlin:1
London:1
2019Conferences:1
```

```
calling:1
papersBulgaria:1
Conference:1
2019User:1
Group:1
EventsSpecial:1
ThanksSocial:1
media@official_phpCopyright:1
©:1
2001-2019:1
GroupMy:1
PHP.netContactOther:1
PHP.net:1
sitesPrivacy:1
policy:1
```



In [ ]:

```
Hence 'PHP' is the most frequent word
```