In [5]:
```
from bs4 import BeautifulSoup
import urllib.request
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /Users/dvora/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

Out[5]: True

In [6]:
```
sr = stopwords.words('english')
print(sr)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "
you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yo
urselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers
', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their',
'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that'
, "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be',
'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did',
'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as',
'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against',
'between', 'into', 'through', 'during', 'before', 'after', 'above', 'b
elow', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', '
under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', '
where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'mo
st', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same
', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don'
, "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 'v
e', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn'
t", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "
haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn'
t", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'was
n', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"
]
```

In [7]:
```
freqWords = dict()
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()]
clean_tokens = tokens[:]
#print(clean_tokens)
for token in tokens:
        if token in stopwords.words('english'):
            clean_tokens.remove(token) #  remove the stop words
```

In [8]:
```
def wordListToFreqDict(wordlist):
    wordfreq = [wordlist.count(p) for p in wordlist]
    return dict(zip(wordlist,wordfreq))
```

In [9]:
```
dictionary = wordListToFreqDict(clean_tokens)
```

In [10]:
```
def sortFreqDict(freqdict):
    sortedDict = [(freqdict[key], key) for key in freqdict]
    sortedDict.sort()
    sortedDict.reverse()
    return sortedDict
```

In [11]:
```
sortFreqDict(dictionary)
```

Out[11]:
```
[(152, 'PHP'),
 (79, 'release'),
 (68, 'found'),
 (32, 'source'),
 (30, 'list'),
 (26, 'also'),
 (25, 'visit'),
 (25, 'team'),
 (25, 'please'),
 (25, 'downloads'),
 (24, 'binaries'),
 (24, 'Windows'),
 (23, 'read'),
 (23, 'ReleasedThe'),
 (23, '7.3.0'),
 (21, 'version'),
 (20, 'next'),
 (20, 'The'),
 (19, 'test'),
 (19, '-'),
 (18, 'us'),
 (18, 'upgrading'),
 (18, 'thebug'),
 (18, 'theUPGRADINGfile'),
 (18, 'theNEWSfile,'),
 (18, 'report'),
 (18, 'planned'),
 (18, 'notes.'),
 (18, 'new'),
 (18, 'make'),
 (18, 'information'),
 (18, 'helping'),
 (18, 'files'),
 (18, 'features'),
 (18, 'complete'),
 (18, 'changes,'),
```

```
(18, 'These'),
(18, 'NOT'),
(18, 'DO'),
(17, 'system.THIS'),
(17, 'sources'),
(17, 'USE'),
(17, 'PRODUCTION!For'),
(17, 'PREVIEW'),
(17, 'IT'),
(17, 'IS'),
(17, 'IN'),
(17, 'DEVELOPMENT'),
(17, 'A'),
(16, 'issues'),
(16, '2018PHP'),
(15, 'thePHP'),
(15, 'changes'),
(15, 'Release'),
(15, '7.2.0'),
(14, 'Candidate'),
(14, '7.3'),
(13, 'would'),
(13, 'thedownload'),
(13, 'specified'),
(13, 'site.Thank'),
(13, 'signatures'),
(13, 'rough'),
(13, 'reporting'),
(13, 'outline'),
(13, 'onthe'),
(13, 'manifestor'),
(13, 'inthe'),
(13, 'glad'),
(13, 'carefully'),
(13, 'archive.The'),
(13, 'announce'),
(13, 'Wiki.For'),
(13, 'QA'),
(12, 'users'),
(12, 'page.'),
(12, 'onwindows.php.net/qa/.Please'),
(12, 'immediate'),
(12, 'encouraged'),
(12, 'development'),
(12, 'cycle'),
(12, 'availability'),
(12, 'announces'),
(12, 'This'),
(12, 'Alpha'),
(9, 'upgrade'),
(8, 'version.For'),
(8, 'theUPGRADING.INTERNALSfile.'),
```

```
(8, 'security'),
(8, 'listed'),
(8, 'Internal'),
(8, '3'),
(7, 'version,'),
(7, 'recorded'),
(7, 'page,'),
(7, 'ourdownloads'),
(7, 'onwindows.php.net/download/.'),
(7, 'Beta'),
(7, 'All'),
(6, 'pre-release,'),
(6, 'bugs'),
(5, 'tracking'),
(5, 'thedownloadpage,'),
(5, 'releases'),
(5, 'incompatibilities'),
(5, 'full'),
(5, 'carefully,'),
(5, 'archive.For'),
(5, 'You'),
(5, 'Aug'),
(5, '2017PHP'),
(4, 'wiki.Thank'),
(4, 'several'),
(4, 'onour'),
(4, 'may'),
(4, 'first'),
(4, 'contains'),
(4, 'bug'),
(4, 'branch.'),
(4, 'atwindows.php.net/qa/.The'),
(4, 'Related'),
(4, 'Jan'),
(4, 'Conference'),
(4, '7.2.0.'),
(4, '7.2'),
(4, '2019PHP'),
(4, '1'),
(3, 'third'),
(3, 'released'),
(3, 'release.'),
(3, 'last'),
(3, 'fixes.All'),
(3, 'final'),
(3, 'alpha'),
(3, 'Sep'),
(3, 'Oct'),
(3, 'Jul'),
(3, 'August'),
(3, '7.0.33'),
(3, '7.0'),
```

```
(3, '5.6.40'),
(3, '5.6'),
(3, '3.'),
(2, 'warrant'),
(2, 'timelines,'),
(2, 'time'),
(2, 'theChangeLog.Please'),
(2, 'theChangeLog.10'),
(2, 'support'),
(2, 'start'),
(2, 'second'),
(2, 'scheduled'),
(2, 'popular'),
(2, 'plans'),
(2, 'otherwise'),
(2, 'one'),
(2, 'note'),
(2, 'menu'),
(2, 'man'),
(2, 'making'),
(2, 'it,'),
(2, 'installation'),
(2, 'important'),
(2, 'good'),
(2, 'fourth'),
(2, 'fixes'),
(2, 'fixed'),
(2, 'discover'),
(2, 'better.28'),
(2, 'beta'),
(2, 'based'),
(2, 'announced'),
(2, 'additional'),
(2, 'according'),
(2, 'There'),
(2, 'Specific'),
(2, 'September'),
(2, 'October.'),
(2, 'October'),
(2, 'November'),
(2, 'Nov'),
(2, 'Jun'),
(2, 'July'),
(2, 'If'),
(2, 'ExtensionsOther'),
(2, 'Dec'),
(2, '7.3.1'),
(2, '7.3.0alpha4'),
(2, '7.3.0RC6'),
(2, '7.3.0RC5'),
(2, '7.3.0RC4'),
(2, '7.3.0RC3'),
```

```
(2, '7.3.0RC2'),
(2, '7.3.0RC1'),
(2, '7.2.2'),
(2, '7.2.14'),
(2, '7.1.26'),
(2, '7.1.25'),
(2, '7.1,'),
(2, '7.1'),
(2, '4'),
(2, '3,'),
(2, '2019Dutch'),
(2, '2.'),
(2, '2,'),
(2, '2'),
(2, '1.'),
(2, '1,'),
(1, '©'),
(1, 'world.Download5.6.40·Release'),
(1, 'websites'),
(1, 'web'),
(1, 'use'),
(1, 'usageGarbage'),
(1, 'uploadsUsing'),
(1, 'tutorialLanguage'),
(1, 'topg'),
(1, 'theChangeLog.22'),
(1, 'theChangeLog.12'),
(1, 'theChangeLog.06'),
(1, 'system.Please'),
(1, 'syntaxTypesVariablesConstantsExpressionsOperatorsControl'),
(1, 'suited'),
(1, 'starts'),
(1, 'sixth'),
(1, 'sitesPrivacy'),
(1, 'sitesMirror'),
(1, 'simple'),
(1, 'seventh'),
(1, 'security-related'),
(1, 'search(current'),
(1, 'search'),
(1, 'scripting'),
(1, 'sGoto'),
(1, 'remote'),
(1, 'release.All'),
(1, 'release,'),
(1, 'relative'),
(1, 'production,'),
(1, 'presumably'),
(1, 'pragmatic,'),
(1, 'powers'),
(1, 'policy'),
(1, 'parametersSupported'),
```

```
(1, 'papersPHPKonf'),
(1, 'pageg'),
(1, 'pageGScroll'),
(1, 'page.Please'),
(1, 'page)/Focus'),
(1, 'pPrevious'),
(1, 'ourwiki.Thank'),
(1, 'options'),
(1, 'open!User'),
(1, 'onwindows.php.net/qa/.The'),
(1, 'nNext'),
(1, 'moduleSession'),
(1, 'minor'),
(1, 'media@official_phpCopyright'),
(1, 'line'),
(1, 'language'),
(1, 'itemkPrevious'),
(1, 'itemg'),
(1, 'included.All'),
(1, 'improvements'),
(1, 'homepageg'),
(1, 'helpjNext'),
(1, 'handlingPersistent'),
(1, 'hGoto'),
(1, 'general-purpose'),
(1, 'gScroll'),
(1, 'flexible'),
(1, 'filesConnection'),
(1, 'file'),
(1, 'fifth'),
(1, 'everything'),
(1, 'especially'),
(1, 'early'),
(1, 'development.Fast,'),
(1, 'cycle,'),
(1, 'considerationsInstalled'),
(1, 'conferencesSunshinePHP'),
(1, 'calling'),
(1, 'bugfix'),
(1, 'boxPHP'),
(1, 'bottomg'),
(1, 'blog'),
(1, 'binaryInstalled'),
(1, 'better.Older'),
(1, 'better.31'),
(1, 'better.30'),
(1, 'better.25'),
(1, 'better.21'),
(1, 'better.19'),
(1, 'better.17'),
(1, 'better.16'),
(1, 'better.13'),
```

```
(1, 'better.11'),
(1, 'better.08'),
(1, 'better.07'),
(1, 'better.06'),
(1, 'better.05'),
(1, 'better.02'),
(1, 'better.01'),
(1, 'authentication'),
(1, 'XFormsHandling'),
(1, 'WrappersSecurityIntroductionGeneral'),
(1, 'VariablesPredefined'),
(1, 'Type'),
(1, 'TracingFunction'),
(1, 'Time'),
(1, 'ThanksSocial'),
(1, 'System'),
(1, 'SupportImage'),
(1, 'Submitted'),
(1, 'StructuresFunctionsClasses'),
(1, 'StartedIntroductionA'),
(1, 'Spring'),
(1, 'Shortcuts?This'),
(1, 'Several'),
(1, 'ServicesWindows'),
(1, 'ServicesSearch'),
(1, 'ServicesCommand'),
(1, 'September.'),
(1, 'SecurityFilesystem'),
(1, 'SecurityError'),
(1, 'SecurityDatabase'),
(1, 'ReportingUsing'),
(1, 'ReleasedPHP'),
(1, 'Register'),
(1, 'ReferenceBasic'),
(1, 'ReferenceAffecting'),
(1, 'RC6,'),
(1, 'RC5,'),
(1, 'RC4.'),
(1, 'RC4,'),
(1, 'RC3.'),
(1, 'RC3,'),
(1, 'RC2,'),
(1, 'RC1,'),
(1, 'QuotesHiding'),
(1, 'Protocols'),
(1, 'ProcessingVariable'),
(1, 'ProcessingCryptography'),
(1, 'Processing'),
(1, 'PreprocessorDownloadsDocumentationGet'),
(1, 'PHPKeeping'),
(1, 'PHPCookiesSessionsDealing'),
(1, 'PHP:'),
```

```
(1, 'PHP.netContactOther'),
(1, 'PHP.net'),
(1, "PHP's"),
(1, 'OutputProcess'),
(1, 'Only'),
(1, 'ObjectsNamespacesErrorsExceptionsGeneratorsReferences'),
(1, 'Notes·Upgrading7.3.1·Release'),
(1, 'Notes·Upgrading7.2.14·Release'),
(1, 'Notes·Upgrading7.1.26·Release'),
(1, 'Notes·Upgrading10'),
(1, 'News'),
(1, 'ModeCommand'),
(1, 'ManipulationGUI'),
(1, 'ManipulationAuthentication'),
(1, 'MIME'),
(1, 'Line'),
(1, 'Language'),
(1, 'June'),
(1, 'July.'),
(1, 'Istanbul'),
(1, 'InvolvedHelpGetting'),
(1, 'Interfaces'),
(1, 'Hypertext'),
(1, 'GroupMy'),
(1, 'Group'),
(1, 'GlobalsUser'),
(1, 'GenerationMail'),
(1, 'Formats'),
(1, 'Five'),
(1, 'Feb'),
(1, 'ExtensionsXML'),
(1, 'ExtensionsWeb'),
(1, 'ExtensionsText'),
(1, 'ExtensionsSession'),
(1, 'ExtensionsServer'),
(1, 'ExtensionsNon-Text'),
(1, 'ExtensionsMathematical'),
(1, 'ExtensionsKeyboard'),
(1, 'ExtensionsHuman'),
(1, 'ExtensionsFile'),
(1, 'ExtensionsDate'),
(1, 'ExtensionsDatabase'),
(1, 'ExtensionsCredit'),
(1, 'ExtensionsCompression'),
(1, 'ExplainedPredefined'),
(1, 'ExceptionsPredefined'),
(1, 'EventsSpecial'),
(1, 'EntriesUpcoming'),
(1, 'Engine'),
(1, 'Encoding'),
(1, 'EditionConferences'),
(1, 'Dynamic'),
```

```
(1, 'December'),
(1, 'Database'),
(1, 'DataMagic'),
(1, 'CurrentFeaturesHTTP'),
(1, 'Control'),
(1, 'ConnectionsSafe'),
(1, 'CollectionDTrace'),
(1, 'ClassesContext'),
(1, 'Character'),
(1, 'CfP'),
(1, 'Card'),
(1, 'CGI'),
(1, 'BehaviourAudio'),
(1, 'Basic'),
(1, 'August.'),
(1, 'Archive'),
(1, 'Apache'),
(1, 'AnnouncementThe'),
(1, '8th.The'),
(1, '7.3.10'),
(1, '7.3.1.'),
(1, '7.3.0beta3.'),
(1, '7.3.0beta3'),
(1, '7.3.0beta2.'),
(1, '7.3.0beta2'),
(1, '7.3.0beta1.'),
(1, '7.3.0beta1'),
(1, '7.3.0alpha4.'),
(1, '7.3.0RC6.'),
(1, '7.3.0RC5.'),
(1, '7.3.0RC4.'),
(1, '7.3.0RC3.'),
(1, '7.3.0RC2.'),
(1, '7.3.0RC1.'),
(1, '7.3.06'),
(1, '7.3.0.beta3'),
(1, '7.3.0.beta2'),
(1, '7.3.0.beta1'),
(1, '7.2.2.'),
(1, '7.2.14.'),
(1, '7.1.26.'),
(1, '7.1.25.'),
(1, '7.0.33.'),
(1, '7.0,'),
(1, '6th.The'),
(1, '5.The'),
(1, '5.6.40.'),
(1, '5.6,'),
(1, '31th'),
(1, '30th.The'),
(1, '2nd.The'),
(1, '27th.The'),
```

```
 (1, '26th'),
 (1, '25th.The'),
 (1, '22nd.The'),
 (1, '21.The'),
 (1, '20th'),
 (1, '2019International'),
 (1, '2019'),
 (1, '2001-2019'),
 (1, '19th.The'),
 (1, '16th.The'),
 (1, '14th'),
 (1, '13th.The'),
 (1, '12th'),
 (1, '11th.The'),
 (1, '(GA),')]
```

In [ ]:
```
import matplotlib.pyplot as plt
%matplotlib inline
plt.figure(2, figsize=(40, 15.2))
freq.plot(100, cumulative=False)
```