In [45]:
```
# Once, download the webtext corpus and import libraries
import re
import json
import nltk
nltk.download('webtext')
```

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

Out[45]: True

In [ ]:
```
# Import the webtext corpus
from nltk.corpus import webtext

# Get the raw text from the corpus, make it lowercase, so using regex is much easier
lines = webtext.raw("firefox.txt").lower()

# Save webtext to a file, to make it easier to work with
with open("firefox.txt", "w", encoding="utf-8") as f:
    f.write(lines)
```

In [47]:
```
# Find all valid links, based on the topdomains.json file, given that every link has a top-level domain and valid url chars
words = []

with open("topdomains.json", "r") as f:
    topdomains = json.load(f)

for topdomain in topdomains:
    words += re.findall(r"\b[\w.:/]+\.{}\b".format(topdomain["fields"]["tld"].lower()), lines)

words = list(set(words))
print(words)
```

```
['osnews.com', 'msn.at', 'gentoo.org', 'pagesjaunes.fr', 'launch.yahoo.com', 'o2.co', 'www.libpr0n.com', 'http://www.cctvusa.com', 'www.
lycos.co.uk', 'news.com', 'microsoft.com', 'mail.yahoo.com', 'eweek.com', 'www.php.net', 'choiceradio.com', 'www.zoneedit.com', 'buy.co
m', 'percentage.it', 'allmakefiles.sh', 'vanguard.com', 'loginnet.passport.com', 'noaa.gov', 'ebay.de', 'yodobashi.com', 'bbc.co.uk', 'w
ww.petetownshend.co.uk', 'slashdot.org', 'www.w3c.org', 'www.google.com', 'cheaptickets.com', 'sidhe.org', 'photo.net', 'theaa.com', 'te
xturizer.net', 'www.blogger.com', 'www.lycos.co', 'client.mk', 'http://labs.google.com', 'www.odeon.co', 'home.pacbell.net', 'koss.com',
'aftonbladet.se', 'www.vipernetworks.com', 'hushmail.com', 'ew.com', 'campusi.com', 'dpreview.com', 'http://www.mozilla.org', 'www.calci
omercato.com', 'www.wamu.com', 'ftp.gimp.org', 'libtoolkitcomps.so', 'www.lexis.com', 'http://www.scripting.com', 'ulasponsel.com', 'ph
p.net', 'nba.com', 'http://ftp.mozilla.org/pub/mozilla.org', 'freebyte.com', 'online.firstusa.com', 'mycard.fleet.com', 'doubleclick.ne
t', 'nsextensionmanager.js.in', 'www.fnac.fr', 'http://www.http://mozilla.org', 'www.rmvplus.de', 'http://www.lexis.com', 'jpbe.net', 'w
ww.microsoft.com', 'gametimesonline.com', 'nu.nl', 'allposters.com', 'http://www.watch.impress.co.jp', 'smartmoney.com', 'libintl.so',
'1.example.com', 'mynetscape.com', 'edmunds.com', 'redirection.pl', 'www.mp3.de', 'net.au', 'www.odeon.co.uk', 'www.file.com', 'sun.co
m', 'www.aol.com', 'login.yahoo.com', 'googlesyndication.com', 'www.logitech.com', 'preprocessor.pl', 'geocaching.com', 'http://www.watc
h.impress.co', 'blogger.com', '2.example.com', 'http://www.trenitalia.com', 'xxx.yyy.com', 'localhost.be', 'gay.com', 'fark.com', 'www.p
etetownshend.co', 'https://www.eposasp.com', 'galleria.net', 'www.x.com', 'verizonwireless.com', 'www.atozwebtools.com', 'https://www.fo
rtify.net', 'yyy.yyyyyy.com', 'www.localhost.net', 'dict.org', 'hku.hk', 'mycheckfree.com', 'mobile.com', 'trease.biz', 'http://texturiz
er.net', 'com.tr', 'iwon.com', 'outwar.com', 'cube.ign.com', 'download.com', 'www.yahoo.com', 'http://www.woolworth.de', 'faltplatte.d
e', 'www.us.army.mil', 'www.uboot.com', 'nppdf.so', 'python.org', 'rottentomatoes.com', 'winamp.com', 'freshmeat.net', 'http://www.odeo
n.co', 'paypal.com', 'libgklayout.so', 'www.alternate.de', 'nationstates.net', 'localhost.net.au', 'libjavaplugin_oji.so', 'google.com',
'www.linuxmail.org', 'koreanair.com', 'http://extensionroom.mozdev.org', 'yahoo.com', 'apple.com', 'adobe.com', 'http://www.timbressuiss
es.ch', 'w.com', 'my.yahoo.com', 'info.org', 'mozilla.sh', 'www.pcpitstop.com', 'o2.co.uk', 'aol.com', 'www.excite.com', 'smartsource.co
m', 'abc.com', 'www.localhost.net.au', 'mozilla.org', 'oo.org', 'pcworld.co', 'folder...in', 'http://www.peterre.com', 'http://www.odeo
n.co.uk', 'fedex.com', 'amazon.com', 'freespeech.org', 'www.xy.com', 'nytimes.com', 'www.intellicast.com', 'localhost.net', 'sequent.or
g', 'usaa.com', 'derstandard.at', 'libprofile.so', 'users.net', 'signonsandiego.com', 'jars.pl', 'foo.com', 'vons.com', 'ford.com', 'ww
w.debian.org', 'suprnova.org', 'bestbuy.com', 'http://bugzilla.mozilla.org', 'vgpro.com', 'nfl.com', 'tylock.com', 'mail.com', 'matchero
o.com', 'www.us', 'moravian.edu', 'www.mozilla.org', 'fstv.org', 'king.fm', 'download.microsoft.com', 'sbc.yahoo.com', 'www.foo.com', 'b
bc.co', 'hottopic.com', 'belinda.ca', 'www.hvv.de', 'mozdev.org', 'libnspr4.so', '2.so', 'pcworld.co.uk', 'mapy.centrum.cz', 'dyndns.or
g', 'sitepoint.com']
```

In [48]:
```
# Find all shortcuts used in the text
shortcut_starters = ["ctrl", "alt", "shift", "cmd", "win", "accel"]

# Find every shortcut in the text, given that it starts with a shortcut starter and has + between letters or numbers or whitespaces
shortcuts = []
for shortcut_starter in shortcut_starters:
    # Find all shortcuts (double or triple) with a + between them
    shortcuts += re.findall(r"(?<!\+\s)(?<!\+\s\s)(?<!\+)\b{}\s*\+\s*\w+\b(?!\s*\+)".format(shortcut_starter), lines)
    shortcuts += re.findall(r"(?<!\+\s)(?<!\+\s\s)(?<!\+)\b{}\s*\+\s*\w+\s*\+\s*\w+\b(?!\s*\+)".format(shortcut_starter), lines)

    # Find all shortcuts (double or triple) with a - between them
    shortcuts += re.findall(r"(?<!\-\s)(?<!\-\s\s)(?<!\-)\b{}\s*\-\s*\w+\b(?!\s*\-)".format(shortcut_starter), lines)
    shortcuts += re.findall(r"(?<!\-\s)(?<!\-\s\s)(?<!\-)\b{}\s*\-\s*\w+\s*\-\s*\w+\b(?!\s*\-)".format(shortcut_starter), lines)

shortcuts = list(set(shortcuts))
print(shortcuts)
```

```
['shift-tab', 'ctrl-a', 'win-m', 'ctrl+t', 'ctrl-click', 'shift + tab', 'cmd-shift-left', 'alt-f4', 'alt+back', 'ctrl+mnemonic', 'alt +
left', 'ctrl+y', 'ctrl+0', 'ctrl + back', 'ctrl-q', 'ctrl+l', 'shift+click', 'ctrl-clicking', 'alt+home', 'alt+scroll', 'alt+f4', 'ctrl+
p', 'ctrl + k', 'ctrl+s', 'shift+enter', 'ctrl+d', 'ctrl-shift-f', 'shift+ctrl+g', 'ctrl +\r\naccessing', 'ctrl+shift', 'ctrl-t', 'alt+f
2', 'ctrl-tab', 'ctrl+backspace', 'ctrl+a', 'ctrl+return', 'alt+f', 'ctrl-y', 'ctrl + view', 'ctrl+mouse', 'ctrl-shift-tab', 'alt-tab',
'ctrl+pgup', 'ctrl- and', 'ctrl+i', 'alt+left', 'ctrl-n', 'alt-enter', 'ctrl+x', 'ctrl+tab', 'ctrl-r', 'alt+click', 'ctrl-down', 'shift
+f6', 'ctrl-middleclick', 'alt + d', 'win-explorer', 'ctrl+ ctrl', 'ctrl+click', 'ctrl+k', 'alt-d', 'ctrl+w', 'alt+f+a', 'cmd-shift-ente
r', 'shift-del', 'shift-doubleclick', 'accel+number', 'accel-shift-l', 'shift + view', 'cmd+m', 'ctrl + shift + l', 'ctrl+b', 'cmd-shift
-h', 'ctrl-x', 'ctrl-up', 'ctrl+f4', 'ctrl-shift-w', 'ctrl+f', 'ctrl+alt+t', 'ctrl-scroll-wheel', 'ctrl+space', 'ctrl+shift+tab', 'cmd-
h', 'accel+shift+g', 'ctrl+q', 'shift+delete', 'shift+link', 'ctrl + middle', 'ctrl-i', 'ctrl-f', 'ctrl+m', 'shift-alt-double', 'ctrl +
enter', 'ctrl-arrow', 'ctrl+mousewheel', 'cmd-enter', 'ctrl+dragging', 'alt+enter', 'ctrl-b', 'ctrl-g', 'shift-f10', 'ctrl-w', 'accel-cl
ick', 'ctrl+e', 'ctrl + b', 'ctrl-m', 'ctrl-k', 'ctrl- scaling', 'ctrl-d', 'ctrl+wheel', 'alt+drag', 'shift+scroll', 'alt + s', 'cmd-cli
ck', 'alt+d', 'alt-home', 'ctrl + f', 'ctrl+enter', 'ctrl + until', 'ctrl-enter']
```