SetupProxyForNutch

Install Tinyproxy

(Ubuntu 11.04 Natty, Kernel Linux 2.6.38-10-generic, GNOME 2.32.1)

Introduction

Tinyproxy is a light-weight HTTP/HTTPS proxy daemon for POSIX operating systems. Designed from the ground up to be fast and yet small, it is an ideal solution for use cases such as embedded deployments where a full featured HTTP proxy is required, but the system resources for a larger proxy are unavailable. Fore more information see here.

Install

```
sudo apt-get install tinyproxy
```

Configure

```
sudo vi /etc/tinyproxy.conf
```

Sample configuration, make sure you set up the Port and Allow (here, I'm using my localhost). **N.B.** Most of these configuration settings are default and can be easily altered to suit. Both the *tinyproxy.conf* configuration file and *LogLevel Info* ensure that the most verbose help is at hand to understand settings and to debug performance.

```
Port 8888
Allow 127.0.0.1
Filter "/etc/filter"
FilterURLs On
FilterDefaultDeny No #filters will act as a blacklist

User nobody
Group nogroup
ViaProxyName "tinyproxy"
ConnectPort 443
ConnectPort 563
Timeout 600
DefaultErrorFile "/usr/share/tinyproxy/default.html"
StatFile "/usr/share/tinyproxy/stats.html"
Logfile "/var/log/tinyproxy/tinyproxy.log"
LogLevel Info
PidFile "/var/run/tinyproxy/tinyproxy.pid"
MaxClients 100
MinSpareServers 5
MaxSpareServers 20
StartServers 10
MaxRequestsPerChild 0
```

Create filters

If necessary these will act as a blacklist, because of FilterDefaultDeny No. This property changes the default policy of the filtering system. If this directive is commented out, or is set to "No" then the default policy is to allow everything which is not specifically denied by the filter file.

However, by setting this directive to "Yes" the default policy becomes to deny everything which is *not* specifically allowed by the filter file e.g. the inverse.

Tinyproxy supports filtering of web sites based on URLs or domains. We need to specify the location of a text file containing the filter rules, one rule per line. This can be done as follows

```
vi /etc/filter
```

and add site urls to be blocked. The list should comprise of single URLs, one per line, just like the seed list for performing crawls.

```
google.com
apache.org
```

for those not experienced using the VI editor please see here for a comprehensive rundown.

Commands to Stop,Start, and Restart

```
sudo /etc/init.d/tinyproxy stop
sudo /etc/init.d/tinyproxy start
sudo /etc/init.d/tinyproxy restart
```

Test the proxy with your browser

For Firefox

- Edit > Preferences > Advanced tab > Network Tab > Connection *Settings* button > select *Manual proxy configuration:* and enter the host you defined above and the port.
- If you have created the filter above, and browse to google.com or apache.org the proxy should block you.

Configure Nutch

Copy the proxy configuration (see below) from conf/nutch-default.xml to conf/nutch-site.xml and fill up with the values of your proxy

```
<property>
  <name>http.proxy.host</name>
  <value>127.0.0.1</value>
  <description>The proxy hostname.  If empty, no proxy is used.</description>
</property>

<property>
  <name>http.proxy.port</name>
  <value>8888</value>
  <description>The proxy port.</description>
</property>
```

Now if you crawl sites, Nutch will use your proxy. You can monitor it by looking at the logs of Tinyproxy during a crawl:

```
sudo tail -f /var/log/tinyproxy.log
```

More resources

- http://ubuntuforums.org/showthread.php?t=122011
- http://en.wikipedia.org/wiki/Tinyproxy
- Tiny Proxy homepage