SetupNutchAndTor

Using Nutch for Crawling Hidden Services (.onion)

Important Note

The aim of this tutorial is to explain **crawling of** hidden services... not for us to use hidden services to crawl. This is a critical point which should both be taken into consideration when reading and using Nutch to crawl the Tor network. Crawling normal websites via Tor can overload the Tor network, but more importantly you can end up making those websites block connections from Tor, thus preventing normal users from being able to reach or use that website. **I f you are looking to use Nutch to crawl the web from behind the Tor network, then you are in the wrong place.**

Introduction

Tor is a network of virtual tunnels that allows people and groups to improve their privacy and security on the Internet. It also enables software developers to create new communication tools with built-in privacy features. Tor provides the foundation for a range of applications that allow organizations and individuals to share information over public networks without compromising their privacy. This tutorial provides an end-to-end example of accessing the Tor network(s) and getting Nutch crawling .onion pages for which the suffix designates an anonymous or pseudonymous address reachable via the Tor network.

Quick Notes

This tutorial has worked best on Debian and Ubuntu however it has also been run on Mac OSX 10.9.4. Best efforts have been made to ensure that documentation covers these OS. If not, then please let us know

Install Tor

This section provides you with three options.

Debian or Ubuntu

Option one is `apt-get install tor` – then you have a socks proxy running on localhost:9050. For many more details, see https://www.torproject.org/docs/debian

If you want to build from source see 1.b below

Mac OSX

Try `brew install tor`, then simply invoke `tor` from the command line. You should see tor starting up, something similar to the following

```
lmcgibbn@LMC-032857 /usr/local/tor(master) $ tor
Sep 23 17:09:47.448 [notice] Tor v0.2.4.23 (git-598c61362f1b3d3e) running on Darwin with Libevent 2.0.21-stable
and OpenSSL 1.0.1i.
Sep 23 17:09:47.448 [notice] Tor can't help you if you use it wrong! Learn how to be safe at https://www.
torproject.org/download/download#warning
Sep 23 17:09:47.449 [notice] Configuration file "/usr/local/etc/tor/torrc" not present, using reasonable
defaults.
Sep 23 17:09:47.452 [notice] Opening Socks listener on 127.0.0.1:9050
Sep 23 17:09:47.000 [notice] Parsing GEOIP IPv4 file /usr/local/Cellar/tor/0.2.4.23_1/share/tor/geoip.
Sep 23 17:09:47.000 [notice] Parsing GEOIP IPv6 file /usr/local/Cellar/tor/0.2.4.23_1/share/tor/geoip6.
Sep 23 17:09:48.000 [notice] Bootstrapped 5%: Connecting to directory server.
Sep 23 17:09:48.000 [notice] Bootstrapped 10%: Finishing handshake with directory server.
Sep 23 17:09:49.000 [notice] Bootstrapped 15%: Establishing an encrypted directory connection.
Sep 23 17:09:49.000 [notice] Bootstrapped 20%: Asking for networkstatus consensus.
Sep 23 17:09:49.000 [notice] Bootstrapped 25%: Loading networkstatus consensus.
...
Sep 23 17:10:12.000 [notice] We now have enough directory information to build circuits.
Sep 23 17:10:12.000 [notice] Bootstrapped 80%: Connecting to the Tor network.
Sep 23 17:10:12.000 [notice] Bootstrapped 90%: Establishing a Tor circuit.
Sep 23 17:10:12.000 [notice] Tor has successfully opened a circuit. Looks like client functionality is working.
Sep 23 17:10:12.000 [notice] Bootstrapped 100%: Done.
```

Cloning Tor Source from Git

```
git clone https://git.torproject.org/git/tor
cd tor
./autogen.sh && ./configure --disable-asciidoc && make
src/or/tor
```

and then you have a socks proxy running on localhost:9050.

Tor Logging

If you want, you can configure your Tor to be more useful in its logging. For example, add these lines to your `/etc/tor/torrc`:

```
SafeLogging 0
LogTimeGranularity 1
```

The Socks Proxy Anomaly

If, as in the case of Nutch, your crawler can't interact with a socks proxy, but it can do an http proxy, then you'll need to run an http proxy and configure it to use a socks proxy. To achieve this we select one of the following proxies.

Polipo

Polipo is a small and fast caching web proxy (a web cache, an HTTP proxy, a proxy server). While Polipo was designed to be used by one person or a small group of people, there is nothing that prevents it from being used by a larger group. You can

On Debian|Ubuntu

```
apt-get install polipo
```

On Mac OSX

```
brew install polipo
```

Then configure your polipo to use Tor: http://www.pps.univ-paris-diderot.fr/~jch/software/polipo/tor.html

That is, set the **socksParentProxy** option in `/etc/polipo/config`

Privoxy

Privoxy is a non-caching web proxy with advanced filtering capabilities for enhancing privacy, modifying web page data and HTTP headers, controlling access, and removing ads and other obnoxious Internet junk. Privoxy has a flexible configuration and can be customized to suit individual needs and tastes. It has application for both stand-alone systems and multi-user networks.

On Debian|Ubuntu

```
apt-get install privoxy
```

On Mac OSX

```
brew install privoxy
```

Then configure your privoxy to use Tor: http://www.privoxy.org/faq/misc.html#TOR

That is, uncomment the **forward-socks5** option in `/etc/privoxy/config` and make sure it points to `127.0.0.1:9050`.

Nutch Crawler Configuration

Configure Nutch to only follow domains that end in .onion. This can be done via simple urlfiltering as described in the main Nutch Tutorial

http://duskgytldkxiuqc6.onion/ is a fine example url to test Nutch on, to make sure you're able to successfully fetch content and metadata.

Then https://ahmia.fi/onions/ has a list of many thousands more, most of which are down so it should be a good exercise for Nutch.

Conclusion

This tutorial acts as a mechanism for using Apache Nutch to crawl hidden services within the Tor network. The intention here is to extend/display/elaborate upon a use case other than typical HTTP protocol crawl cycles. Hopefully this tutorial provides that. The most important thing here is for people to maintain this docuementation. If there is something which does not work, then please let us know