Dr. Scrapelove

(or: How I Learned to Beat Anti-Scrape Websites and Love WWW::Mechanize::Firefox)

By Trevor Cordes

MUUG Presentation June 2014

Legal Disclaimer

- Check site usage policies
- Legally or contractually: may be prohibited
- Ethically: OK
- Scraping = Hacker != Cracker
- Scraping = accessing content legally available to you, but at a faster speed, and providing you with a copy, all automated
- Be nice: sleep(rand(10))

What's The Use?

- Transform into a more useful format
- Impermanent web data
- Time-limited web data (i.e. subscriptions)
- For-pay newspaper sites, consumerreports.org
- Automate web tasks
- Yes, pr0n

What I've Done

- Programmed scrapers for customers: Scraped entire sites for content
- Automated a few daily/weekly internet chores
- My hobby: Scraped a certain web auction site for 11 years tracking every sale of about 5000 audio CD titles, and bought a copy of each at or near the record lowest price. Sell some when prices are high.

Demo

Simple FF Remote

- firefox -remote 'openURL($url)' >/dev/null 2>&1 &
- Opens $url on the most-recently-clicked FF window
- Simple
- Can use from shell command line or any programming language

Scraping: Different Levels

- wget -r http://muug.mb.ca
- WWW::Mechanize
- Perl, also Python, Ruby, etc
- WWW::Mechanize::Firefox with MozRepl
- The Future: Mechs with Javascript engines

WWW::Mechanize

- Perl module
- Pretends to be a browser
- OO interface
- Easy, Fast
- Authenticate with obsolete http basic auth and simple non-javascript login systems
- Getting rarer
- Install via package manager: perl-WWW-Mechanize on Fedora

“Web Developer”

- Handy little Firefox add-on
- Install in usual manner
- Provides a “view generated source” function
- Firefox's “View Source” is near useless
- WD's generated-source option reflects what you actually see on screen
- Post-Javascript, post-AJAX, post-CSS, etc
- Also good, CONTROL-SHIFT-C, needed for IFrames

Demo

- 0-ebay: Scrape some data off ebay
- 1-pcplus: Oops, foiled!

*@&(! Javascript

- Many site logins require Javascript
- WWW::Mechanize has no JS engine
- Workarounds, wireshark, mentally parsing .js
- Some sites obfuscate login/session, js hashes
- Engine In the works
  - Perhaps in the future
  - Target a browser?

WWW::Mechanize::Firefox

- Instead of perl-as-browser...
- Uses Firefox as the browser
- Perl simply is the remote control
- Uses MozRepl

MozRepl

- MozRepl Firefox add-on
- Provides a telnet interface into your actual, running, browser. Cool!
- Can get data, set data, control functions
- Install:
    - Doesn't appear in add-on search, so use:
    - https://addons.mozilla.org/en-US/firefox/addon/mozrepl/
    - afterwards, press F10 to see menu bar
    - then Tools->MozRepl->start
    - and also activate-on-start if desired

MozRepl Demo

- telnet localhost 4242
- window.alert("Hi MUUGers")
- document.title
- document.title="MUUGers Window"
- content.location.href='http://muug.mb.ca'
- repl.quit()

WWW::Mechanize::Firefox

- Bleeding edge = Difficult install
- Via rpms on Fedora:
  - perl-HTML-Selector-XPath
  - perl-IPC-Run
  - perl-JSON
  - perl-Carp-Clan
  - perl-Class-Accessor
  - perl-Class-Data-Inheritable
  - perl-Data-Dump
  - perl-Net-Telnet
  - perl-Template-Toolkit
  - perl-Text-SimpleTable
  - perl-UNIVERSAL-require
  - perl-Params-Util
  - perl-MRO-Compat
  - perl-LWP-Protocol-https

- Via CPAN:
  - Class::Default
  - Data::JavaScript::Anon
  - Module::Pluggable::Fast
  - Template::Provider::FromDATA
  - MozRepl
  - MozRepl::RemoteObject
  - Object::Import
  - Shell::Command
  - WWW::Mechanize::Firefox
- example: cpan
- or perl -MCPAN -e shell
- install Class::Default
- disable follow!

Demo: pcplus

- Who wants to manually login every week when they ring their Pavlovian bell?
- Can't we automate?
- Site requires Javascript for login
- Demo: 3-pcplus-firefox

Or Is It Web Scraping?

- Wikipedia says: “Screen scraping is normally associated with the programmatic collection of visual data from a source, instead of parsing data as in web scraping.”
- Trevor's Screen Scraping Definition: Getting what you want of theirs onto yours, and not taking no for an answer.