Scrape the Web

Strategies for programming websites that don't expect it

Asheesh Laroia asheesh@asheesh.org

January 18, 2012

This is supposed to be practice

- But there are things I know I’ll have to fix

This is supposed to be practice

- But there are things I know I'll have to fix
- More graphs and diagrams

This is supposed to be practice

- But there are things I know I'll have to fix
- More graphs and diagrams
  - But which ones?

This is supposed to be practice

- But there are things I know I'll have to fix
- More graphs and diagrams
  - But which ones?
- Better presentation style (sorry!)

So please do interrupt me

...and we begin

Welcome

Meta

- You will learn neat tricks

Meta

- You will learn neat tricks
- DO NOT BECOME AN EVIL COMMENT SPAMMER

Meta

- You will learn neat tricks
- DO NOT BECOME AN EVIL COMMENT SPAMMER
- Theory and running code

Meta

- You will learn neat tricks
- DO NOT BECOME AN EVIL COMMENT SPAMMER
- Theory and running code
- Brittle? Sometimes.

Meta

- You will learn neat tricks
- DO NOT BECOME AN EVIL COMMENT SPAMMER
- Theory and running code
- Brittle? Sometimes.
- The comics aren't mine; ask for references.

Things you'll need

- Sample code: http://FIXME.com
- Install FireBug
- Install python-lxml if it is easy

Pacing

- Slow me down,

Pacing

- Slow me down,
- or speed me up.
- With your voice, or by raising your hand.

What is web scraping?

Generally speaking,

- You retrieve some data from the web,

What is web scraping?

Generally speaking,

- You retrieve some data from the web,
- You extract some information,

What is web scraping?

Generally speaking,

- You retrieve some data from the web,
- You extract some information,
- and optionally you repeat.

Perspectives on scraping

- One page vs. a whole site

Perspectives on scraping

- One page vs. a whole site
- A site's contents now, or for the future as well

(1) Diving in with curry

Lunchtime

- http://mehfilindian.com/LunchMenuTakeOut.htm

Lunchtime

- http://mehfilindian.com/LunchMenuTakeOut.htm
- A question

Lunchtime

- http://mehfilindian.com/LunchMenuTakeOut.htm
- A question
  - is there eggplant today?

From python

examples/curry/trivial.py

(2) HTML: Structured text on the web

Two easy ways to read HTML

In a browser:

- View source

Two easy ways to read HTML

In a browser:

- View source
- Inspect element (requires Firebug or DOM Inspector)

HTML vs. XHTML

- Both are trees of tags

HTML vs. XHTML

- Both are trees of tags
- HTML: from 1992

HTML vs. XHTML

- Both are trees of tags
- HTML: from 1992
- XHTML: from 2000

HTML vs. XHTML

- Both are trees of tags
- HTML: from 1992
- XHTML: from 2000
- ...did XHTML win?

Stats pop quiz: size and type

(Stats from the MAMA survey published by Opera, http://dev.opera.com/articles/view/mama-key-findings/.)

- Average page size?

Stats pop quiz: size and type

(Stats from the MAMA survey published by Opera, http://dev.opera.com/articles/view/mama-key-findings/.)

- Average page size?
  - 16.5K

Stats pop quiz: size and type

(Stats from the MAMA survey published by Opera, http://dev.opera.com/articles/view/mama-key-findings/.)

- Average page size?
    - 16.5K
- HTML to XHTML ratio?

Stats pop quiz: size and type

(Stats from the MAMA survey published by Opera, http://dev.opera.com/articles/view/mama-key-findings/.)

- Average page size?
  - 16.5K
- HTML to XHTML ratio?
  - 2:1

Stats pop quiz: quirks and tags (pt. 1)

- Transitional vs. Strict+Framset?

Stats pop quiz: quirks and tags (pt. 1)

- Transitional vs. Strict+Framset?
  - 10:1

Stats pop quiz: quirks and tags (pt. 1)

- Transitional vs. Strict+Framset?
  - 10:1
- How many pages render in "Quirks" mode?

Stats pop quiz: quirks and tags (pt. 1)

- Transitional vs. Strict+Framset?
  - 10:1
- How many pages render in "Quirks" mode?
  - 85 percent

Stats pop quiz: quirks and tags (pt. 1)

- Transitional vs. Strict+Framset?
  - 10:1
- How many pages render in "Quirks" mode?
  - 85 percent
- What percent validate?

Stats pop quiz: quirks and tags (pt. 1)

- Transitional vs. Strict+Framset?
  - 10:1
- How many pages render in "Quirks" mode?
  - 85 percent
- What percent validate?
  - 4.13 percent

Stats pop quiz: quirks and tags (pt. 2)

- What's more popular? TITLE or BODY?
- The web is a mess

Stats pop quiz: quirks and tags (pt. 2)

- What's more popular? TITLE or BODY?
  - TITLE
- The web is a mess

Stats pop quiz: quirks and tags (pt. 2)

- What's more popular? TITLE or BODY?
  - TITLE
- What percent of web pages with validations badges actually validate?

- The web is a mess

Stats pop quiz: quirks and tags (pt. 2)

- What's more popular? TITLE or BODY?
  - TITLE
- What percent of web pages with validations badges actually validate?
  - about 50 percent
- The web is a mess

(3) Parsing "HTML" in Python

Things to consider

- What does a parser do with invalid HTML?

Things to consider

- What does a parser do with invalid HTML?
- Does it handle XHTML properly?

Things to consider

- What does a parser do with invalid HTML?
- Does it handle XHTML properly?
    - They all do; don't worry.

Things to consider

- What does a parser do with invalid HTML?
- Does it handle XHTML properly?
  - They all do; don't worry.
- examples/parsing/ has samples.

A showcase of some options

- HTMLParser (stdlib!)
- xml.dom.minidom (stdlib!)

A showcase of some options

- HTMLParser (stdlib!)
- xml.dom.minidom (stdlib!)
- BeautifulSoup
- html5lib
- lxml.html

(4) On regular expressions

- Some people, when confronted with a problem, think
    - "I know, I'll use regular expressions."
- Now they have two problems. – jwz.

But why?

- a href=”whatever”
- a href=’whatever’
- a href=”whatever’

But it's good enough for...

- Curry

But it's good enough for...

- Curry
- Text analysis:

But it's good enough for...

- Curry
- Text analysis:
    - Reviews 1-10 of 430

If you have to

- Use Kodos, a regular expression GUI
- (Note: redemo.py in Python source is unmaintained.)

If you have to

- Use Kodos, a regular expression GUI
- (Note: redemo.py in Python source is unmaintained.)
- Be conservative in what you do, be liberal in what you accept from others. – Jon Postel.

(5) Parsers in depth

Searching document trees

- BeautifulSoup API (examples/tree-builders/beautifulsoup/search.py)

Searching document trees

- BeautifulSoup API (examples/tree-builders/beautifulsoup/search.py)
- html5lib creates BeautifulSoup objects (or others) (examples/tree-builders/html5lib/search.py)

Searching document trees

- BeautifulSoup API (examples/tree-builders/beautifulsoup/search.py)
- html5lib creates BeautifulSoup objects (or others) (examples/tree-builders/html5lib/search.py)
- lxml provides XPath (examples/tree-builders/lxml/search xpath.py)

Searching document trees

- BeautifulSoup API (examples/tree-builders/beautifulsoup/search.py)
- html5lib creates BeautifulSoup objects (or others) (examples/tree-builders/html5lib/search.py)
- lxml provides XPath (examples/tree-builders/lxml/search xpath.py)
- lxml provides CSSSelect (examples/tree-builders/lxml/search css.py)

General structure of scraping a page

- Get a page's HTML

General structure of scraping a page

- Get a page's HTML
- Parse it

General structure of scraping a page

- Get a page's HTML
- Parse it
- Pull out the items you need

General structure of scraping a page

- Get a page's HTML
- Parse it
- Pull out the items you need
- Return them as a dictionary, or an object

A closer look at curry

- (see Python interpreter)
- (let's use BeautifulSoup)

A closer look at curry

- (see Python interpreter)
- (let's use BeautifulSoup)
- Conclusion: This is a text-processing problem, not a tag problem.

Mini-lesson

Three kinds of page:

- Hand-written pages

Mini-lesson

Three kinds of page:

- Hand-written pages
- Machine-written pages

Mini-lesson

Three kinds of page:

- Hand-written pages
- Machine-written pages
- Machine-written pages, old-skool

More BeautifulSoup: Yahoo! Finance

- examples/tree-builders/beautifulsoup yfinance.py

(6) Interacting with the web

Hard-coding URLs: Yahoo! search

- examples/search/yahoo.py
- examples/search/google.py

More about HTTP: Headers

- (Firefox demo)

More about HTTP: Status codes

- 2xx: Success
- 3xx: Redirection
- 4xx: Error

More about HTTP: Status codes

- 2xx: Success
- 3xx: Redirection
- 4xx: Error
- 402: Payment Required
- 404 Not Found

More about HTTP: Status codes

- 2xx: Success
- 3xx: Redirection
- 4xx: Error
- 402: Payment Required
- 404 Not Found
- 410 Gone

More about HTTP: Status codes

- 2xx: Success
- 3xx: Redirection
- 4xx: Error
- 402: Payment Required
- 404 Not Found
- 410 Gone
- 418 I'm a teapot

More about HTTP: Methods

- GET
- POST
- PUT

More about HTTP: Methods

- GET
- POST
- PUT
- BREW

Once we set User-Agent, are we just like Firefox?

- JavaScript behavior

Once we set User-Agent, are we just like Firefox?

- JavaScript behavior
- Image download behavior

Once we set User-Agent, are we just like Firefox?

- JavaScript behavior
- Image download behavior
- Cookie behavior

Once we set User-Agent, are we just like Firefox?

- JavaScript behavior
- Image download behavior
- Cookie behavior
- Invalid HTML handling behavior (?)

Once we set User-Agent, are we just like Firefox?

- JavaScript behavior
- Image download behavior
- Cookie behavior
- Invalid HTML handling behavior (?)
- Accept: headers

Google, again

- examples/search/urllib2-user-agent/google as ie.py

Google, again

- examples/search/urllib2-user-agent/google as ie.py
- IE 5 vs. IE 8

robots.txt

- User-agent: *
- Disallow: /
- Allow: /crawlme.html

robots.txt

- User-agent: *
- Disallow: /
- Allow: /crawlme.html
- http://www.robotstxt.org/

robots.txt

- User-agent: *
- Disallow: /
- Allow: /crawlme.html
- http://www.robotstxt.org/
- Don't ever GET it

(7) Filling out forms, and handling cookies, with mechanize

The weather (by hand)

- http://cepstral.com/cgi-bin/demos/weather

The weather (by hand)

- http://cepstral.com/cgi-bin/demos/weather
- Find the POST target in Firebug

The weather (by hand)

- http://cepstral.com/cgi-bin/demos/weather
- Find the POST target in Firebug
- examples/cepstral/just post.py

The weather (by hand)

- http://cepstral.com/cgi-bin/demos/weather
- Find the POST target in Firebug
- examples/cepstral/just post.py
- examples/cepstral/play wav.py

The weather (with mechanize)

- examples/cepstral/just post via mechanize.py

Search the web (via mechanize)

- examples/search/yahoo mechanize.py

Search the web (via mechanize)

- examples/search/yahoo mechanize.py
- Oh snap, we're a robot.

Search the web (via mechanize)

- examples/search/yahoo mechanize.py
- Oh snap, we're a robot.
- examples/search/yahoo mechanize norobots.py

Search the web (via mechanize)

- examples/search/yahoo mechanize.py
- Oh snap, we're a robot.
- examples/search/yahoo mechanize norobots.py
- examples/search/google mechanize.py

(8) Recap and philosophy

Things we've seen

- Loading web pages from the network with urllib2

Things we've seen

- Loading web pages from the network with urllib2
- Parsing web pages (even broken ones)

Things we've seen

- Loading web pages from the network with urllib2
- Parsing web pages (even broken ones)
- Scraping that page into a set of structured Python objects

Things we've seen

- Loading web pages from the network with urllib2
- Parsing web pages (even broken ones)
- Scraping that page into a set of structured Python objects
- HTTP status codes

Things we've seen

- Loading web pages from the network with urllib2
- Parsing web pages (even broken ones)
- Scraping that page into a set of structured Python objects
- HTTP status codes
- Faking the user agent header

Things we've seen

- Loading web pages from the network with urllib2
- Parsing web pages (even broken ones)
- Scraping that page into a set of structured Python objects
- HTTP status codes
- Faking the user agent header
- Submitting forms

Things we've seen

- Loading web pages from the network with urllib2
- Parsing web pages (even broken ones)
- Scraping that page into a set of structured Python objects
- HTTP status codes
- Faking the user agent header
- Submitting forms
- Keeping a session with cookies

(9) Even more about parsers

Things to look for

- Performance
- Ease-of-use
- Quality
- Maintained-ness

Redux

- HTMLParser (stdlib!)

Redux

- xml.dom.minidom (stdlib!)

Redux

- xml.dom.minidom (stdlib!)
- BeautifulSoup

Redux

- xml.dom.minidom (stdlib!)
- BeautifulSoup
- html5lib

Redux

- xml.dom.minidom (stdlib!)
- BeautifulSoup
- html5lib
- lxml.html

Winners

- Resilience: lxml.html == html5lib ¿ BeautifulSoup ¿ stdlib

Winners

- Resilience: lxml.html == html5lib ¿ BeautifulSoup ¿ stdlib
- Performance: lxml.html ¿ stdlib ¿ BeautifulSoup ¿ html5lib

(10) Countermeasures

The basics

- Referer header
- Cookies
- Hidden form fields

The basics

- Referer header
- Cookies
- Hidden form fields
- Solved by mechanize

The hard ones

- Per-IP address query limits

The hard ones

- Per-IP address query limits
- Behavior profiling

The hard ones

- Per-IP address query limits
- Behavior profiling
- JavaScript

The hard ones

- Per-IP address query limits
- Behavior profiling
- JavaScript
- CAPTCHAs

IP addresss

- ssh -D
- tsocks
- socks monkey

IP addresss

- ssh -D
- tsocks
- socks monkey
- (All in the sample code)

Behavior profiling

- You're doomed.

JavaScript

Options:

1. Re-write the JS in Python
2. Send the JS to python-spidermonkey

CAPTCHAs

- Hope for an easy one
  - http://www.mailinator.com/images/captcha1.gif

CAPTCHAs

- Hope for an easy one
  - http://www.mailinator.com/images/captcha1.gif
- Otherwise, just present it to the human operator...

(11) When (and how) to just automate Firefox

Selenium and friends

- examples/seleniumrc/test google.py
- Selenium RC for XPath

Thanks

asheesh@asheesh.org