

Identifying areas of interest when web scraping

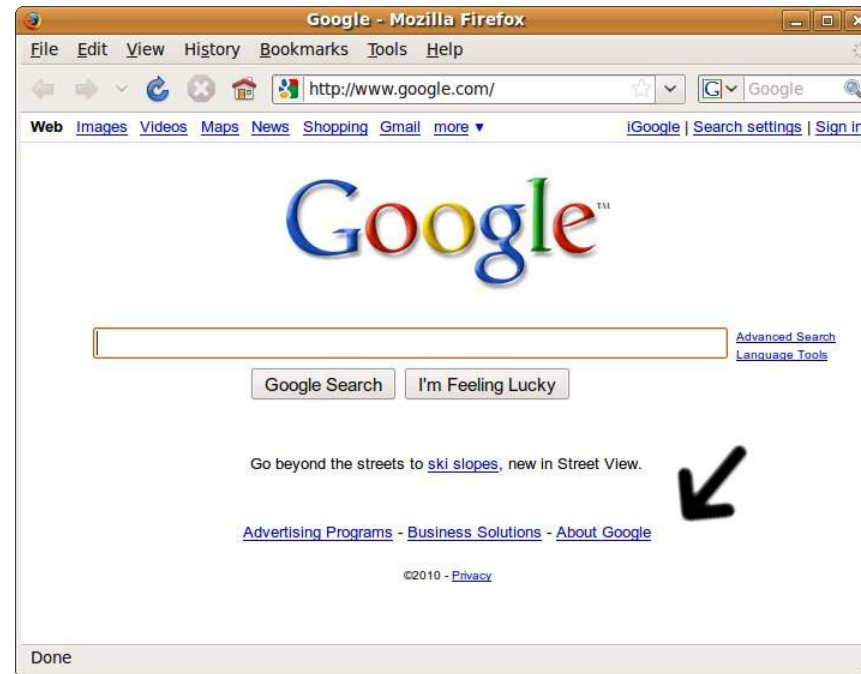
Presented to the Southern California Python Interest Group 2/11/2010

Mel Raab mel.r.01@gmail.com

Web Scraping: gather posted information not typically provided in a more computer-friendly form

Typically requires encoding human understanding of the location of information on a page.

*(Example:
find link
reference)*



Search source and identify About Google

```
<a href="/intl/en/about.html">About Google</a>
```

Rigid encoding can make for a brittle solution

- what if the text prompt is changed?
- what if another link is added, changing the index?

Sometimes the search is not well focused

- example: visit CNN, search for Super Bowl section



Super Bowl »

- [Saints capture first Super Bowl title](#)
- [High-tech kept Super Bowl on track](#)
- [Opinion: Curse of the instant replay](#)
- [Best, worst Super Bowl ads Time](#)
- [Team's success rubs off on city](#)

Search source and identify Super Bowl

Betty White, 88, is red-hot right now with a popular **Super Bowl...**

Kendra: Why I cried after **Super Bowl ...**

-most-watched-show-ever/?hpt=Sbin">Marquee: **Super Bowl ...**

ads-topped-super-bowl/?hpt=Sbin">**Super Bowl** ads

Super Bowl

Need a better way to locate information of interest.

Information for human visual presentation
is usually chunked with structure.

Use page structure to help identify areas of interest.

Example direct use of page structure

“from top of page, go down 3 divs”

“locate the next table”

“go to the 3rd row of table”

“go to the 2nd column of the table”

(“extract and save information”)

- This approach assumes static page layout
- Useless if chunk is moved to another part of page
- Relies on inflexible absolute addressing
- Subject to failure with imperfect HTML

Scraping structured text

...from Python in a Nutshell

“Python supplies `sgmlib`, `htmlib` and `HTMLParser`”

These traverse a page, tag by tag, and let you take action on a per-tag basis. No interest or support for page structure.

...from Python in a Nutshell

“When ... dealing with broken web pages, ... `BeautifulSoup` offers your best, last hope.”

Beautiful Soup

<http://www.crummy.com/software/BeautifulSoup/>

- Python HTML/XML parser
- “Won't choke if you give it bad markup.”
- “Provides a few simple methods and Pythonic idioms for navigating, searching”
- “Parses anything you give it”
- “Find all the links”
- “Find all the links of class externalLink”
- “Find all the links whose urls match 'foo.com' ”
- “Find the table heading that's got bold text, then give me that text.”

Used by **Mechanize**, adds browser function

<http://wwwsearch.sourceforge.net/mechanize/>

Unfortunately, Beautiful Soup does choke.

Slashdot

```
<a href="//slashdot.org">
```

...and on this actual page that renders in browsers:

```
<link href="...">
```

```
<HTML>
```

```
<HEAD>
```

```
</head>
```

```
<BODY ...>
```

```
<table ...>
```

```
...
```

```
<html>
```

```
<head>
```

```
<meta ...>
```

```
<title>
```

```
</title>
```

```
</head>
```

```
<body>
```

```
<table ...>
```

```
...
```

```
</table>
```

```
</body>
```

```
</html>
```

```
...
```

```
<html>
```

```
<head>
```

```
</head>
```

```
<body>
```

```
...
```

```
</body>
```

```
</html>
```

```
</td>
```

```
</tr>
```

```
</table>
```

```
</body>
```

```
</html>
```

Instead, recognize page structure can be discerned from a sequence of HTML tags and their payloads

```
<div class="cnn_sectbin3">
<div class="cnn_sectbincontnt">
<div class="cnn_sectbincontnt2">
<h4>
<a href="/SPECIALS/2010/super.bowl/?hpt=Sbin">Super Bowl
</a>
</h4>

<div class="cnn_clear">
</div>
<div class="cnn_divline">
</div>
<ul class="cnn_bulletbin">
<li>
<a href="/2010/SPORT/02/07/...?hpt=Sbin">Saints capture
first Super Bowl title
</a>
</li>
```

Look at the pattern of HTML tags

```
div
div
div
h4
a
/a
/h4
```

(manually decide that to here is sufficient for chunk identification;
not a limitation of the technique)

```
div
/div
div
/div
ul
li
a
/a
/li
```

Replace tags with tokens of uniform size

```
HTMLtoToken = {  
  HTMLtoToken[ '<!-- ' ]      = 'C0 '  
  HTMLtoToken[ '-->' ]      = 'C1 '  
  HTMLtoToken[ 'a' ]         = 'A0 '  
  HTMLtoToken[ '/a' ]        = 'A1 '  
  HTMLtoToken[ 'abbr' ]      = 'A2 '  
  HTMLtoToken[ '/abbr' ]     = 'A3 '  
  HTMLtoToken[ 'acronym' ]   = 'A4 '  
  HTMLtoToken[ '/acronym' ] = 'A5 '  
  HTMLtoToken[ 'address' ]   = 'A6 '  
  HTMLtoToken[ '/address' ] = 'A7 '  
  HTMLtoToken[ 'applet' ]    = 'A8 '  
  HTMLtoToken[ '/applet' ]   = 'A9 '  
  HTMLtoToken[ 'area' ]      = 'a0 '  
  HTMLtoToken[ 'b' ]         = 'B0 '  
  HTMLtoToken[ '/b' ]        = 'B1 '  
  HTMLtoToken[ 'base' ]      = 'B2 '  
  HTMLtoToken[ 'basefont' ] = 'B3 '  
}
```

In our CNN example ...

`div` → `D8`
`div` → `D8`
`div` → `D8`
`h4` → `H4`
`a` → `A0`
`/a` → `A1`
`/h4` → `h4`
`div` → `D8`
`/div` → `D9`
`div` → `D8`
`/div` → `D9`
`ul` → `U2`
`li` → `L4`
`a` → `A0`
`/a` → `A1`
`/li` → `L5`

(decided that to here is sufficient for chunk identification)

`D8D8D8H4A0A1h4D8D9D8D9U2L4A0A1L5`

Can use `string.find()` to locate

The pertinent token sequence is

D8D8D8H4A0A1h4

Let's have a look at the CNN web page as it renders in Firefox.

Note how the chunks “U.S.”, “World”, “Business” are grouped together.

Next come “Politics”, “Entertainment” and “Health.”

Afterwards, seven chunks are grouped together.

updated 2:46 p.m. EST, Tue February 9, 2010



Obama open to small steps on job growth

President Obama today said he had "a good and frank conversation" with a bipartisan delegation of lawmakers on new initiatives to spark job growth. FULL STORY

• Can jobs bring parties together?

Latest news

- Dow surges 190 points CNNMoney
- 10-20 inches of snow forecast for D.C.
- Travelers face canceled flights, delays
- Prius recalled? What to do CNNMoney
- Timeline of Toyota troubles CNNMoney
- Mrs. Obama to kids: Let's Move
- Iran ramps up uranium enrichment
- Dr. Gupta: Why I went back to Haiti
- D.A.: Fight killed Kerrigan's dad WCVB
- 60 killed in Afghanistan avalanche
- Ticker: Gov. denies sex, drug rumors
- Ron Reagan blasts Sarah Palin



Cool Betty White is red-hot at 88

Veteran actress Betty White, 88, is red-hot right now with a popular Super Bowl ad and a campaign to get her as a host for "Saturday Night Live." Here's a look at a not-so-overnight success. FULL STORY

DON'T MISS



Joe Jackson: Charge 'not enough' 6:07



Haitian found after 4 weeks in rubble



Life without a limb 2:18



The fight to end child sex slavery



Palin palm message mocks critics



Boy Scouts turn 100

Ford Escape
Escape quality can't be beat by Toyota Rav4.
Learn More

Build & Price Colors and 360 Photos Features

ADVERTISEMENT

Hi! Log in or sign up to get great personalized stuff here!

NEWSPULSE LOCAL WEATHER & NEWS SPORTS MARKETS

Updated 2:50 pm ET Feb 9

Dow 10,073.07 +164.68 (+1.66%)

Nasdaq 2,151.22 +25.17 (+1.18%)

S&P 1,070.75 +14.01 (+1.33%)

Enter Symbol Get quotes

- Ron Reagan blasts Sarah Palin
 - Airline to charge \$8 for blankets
 - 500 homes ordered evacuated
 - 100 animals die in crash
 - 10 Texas churches burned
 - Google diving deeper into social media?
 - Alba 'distressed' by lookalike plans
 - Black belt takes down unruly flier
- [View more stories](#)

The fight to end child sex slavery



Sandra Bullock downplays Oscar bid

Palin palm message mocks critics



Dubai diners flock to eat 'camel burger'

Boy Scouts turn 100



Insider's guide to attending Olympics

Enter Symbol

[Get quotes](#)

Editor's choice Top picks Must-watch video




Actress selling home at \$2M loss



Sanford oblivious to pain, wife recalls



Fit Nation: Weight loss tips



The People vs. George Lucas



Drive one.

ADVERTISEMENT



Matters of the Heart: Remedy risk



Track jobs in your state

U.S. »

- See your snowy mid-Atlantic photos
- Reagan vs. Palin
- Prius owner stands by his car
- Deer into motel
- Kids asked to solve world problems
- Sanford: Memoir 'cathartic'
- 1000+ fall ill in mumps outbreak

[More](#)

World »

- Girl victim of Cambodia's sex trade
- Saving Africa's white lion
- Iran ramps up enrichment
- UK Afghan casualties reach grim mark
- Camel burgers hot in Dubai
- Climate chief pens steamy novel
- World's tallest viewing deck closed

[More](#)

Business »

- Unemployment taxes slam biz
- Tax banker bonuses? Yeah, right
- The 401(k) match is back!
- A foreclosure money pit
- Scarlett Johansson's housing woes
- Double-dip recession protection
- Toyota workers talk about fix

[More](#)

Quick vote

Would you pay \$8 for a blanket and pillow during airline travel?

Yes No

[VOTE](#) or [view results](#)

SPONSORED BY **EQUIFAX**

Politics »

- Health care debate goes live
- Analysis: Now what, Tea Party?

Entertainment »

- 'Bachelor' wasn't 'awesome'
- Cool Betty White is red-hot

Health »

- First lady takes on childhood obesity
- Surgery error, Murtha death linked

Sponsored links

Valentine's Day Jewelry
Save 40-60% On Select Fine Jewelry At JCPenney. Savings End 2/10.

- Analysis: Now what, Tea Party?
- Budget hearings: Policy, presentation
- Palin: 2012 presidential bid possible
- Mayor Bernero announces gov run
- WWII-era navigation system closing
- Pennsylvania Dems back Specter

More

- Cool Betty White is red-hot
- Kate Gosselin to release new book
- Kendra: Why I cried after Super Bowl
- Jeff Probst renews 'Survivor' contract
- Bullock downplays Oscar bid
- No longer worried about '24'

More

- Surgery error, Murtha death linked
- Gupta: Why I'm back in Haiti
- Talking to your kids about healthy diet
- Study: 'E-cigarettes' don't deliver
- Older moms boost autism risk
- H1N1 flu continues to kill, CDC warns

More

JCPenney. Savings End 2/10. JCPenney.com

Are you "PM" Certified? Villanova Project Management Certification 8 weeks - Enroll Now. www.VillanovaU.com

LendingTree Official Site Your loan, Your options. Take Control at the all new LendingTree® www.LendingTree.com

Tech »

- New Apple product is ... Aperture 3
- Google making dive into social media?
- Engineer's quest to caption the Web
- Facebook gets birthday face-lift
- China: Hacker training site shut down
- 'Off grid' brings power to the people
- Google: U.S. needs faster Internet

More

Living »

- When is the ideal time to get married?
- Physical contact at work -- watch out
- She howled for help on Twitter
- Interview the 'love of your life'
- Be sane, surrender to the clutter
- They live in a time warp -- and enjoy it!
- He paid off her loans, she feels guilty

More

Justice »

- Joe Jackson: Dr. Murray 'a fall guy'
- What is involuntary manslaughter?
- Police: Child tortured over ABC's
- Exonerated man, accuser forge bond
- Cops: U.S. missionaries tried before
- Boy slain in 1990; brother still missing
- Judge: Turn over Edwards sex tape

More

CNN Challenge »



Take our new online news trivia quiz hosted by your favorite CNN anchors.

Play

Sports »

- Best skater in decades a loner
- Swimsuit 2010: Olympic athletes
- What's next for Saints, Colts?
- GM: No contract talks with Jeter, Mo
- Celebrities pick Daytona 500 winner
- Big decision pending on superstar
- SI's memorable photos of the week

More

Blogs »

- Marquee: Super Bowl owns TV record
- AC360°: New Orleans inspires
- amFix: Super Bowl ads
- Afghan Crossroads: War secret
- amFix: Pioneering paparazzo
- amFix: Palin and the Tea Party
- Marquee: Lil Wayne to be sentenced

More

Super Bowl »

- Saints capture first Super Bowl title
- High-tech kept Super Bowl on track
- Opinion: Curse of the instant replay
- Best, worst Super Bowl ads Time
- Team's success rubs off on city
- Gallery: Halftime highlights SI
- Two unknowns keyed Saints victory SI

More

Hot topics »

- 1 Winter Weather
- 2 Toyota Recall
- 3 Michael Jackson
- 4 Air Travel
- 5 Obesity
- 6 Super Bowl
- 7 New Orleans Saints

Summary

- manually identify visual chunk of interest
- discern identifiable HTML structure of chunk
- express structure of chunk with uniform length tokens
- express structure of entire page with uniform length tokens
- find() expression of chunk in expression of page
- deeper text pattern search tools can investigate structure
- use index of chunk to point back to representation of original page source
- continue gathering information from identified location
- chunk may be relocated in page and still found
- chunk absence can trigger exception
- page structure doesn't have to be normalized first, technique successful even if structure doesn't make sense