

Python Scraping Showdown

A speed and accuracy comparison

Katharine Jarmul (@kjam)

PyCon 2014

About the Speaker

- Been using scrapers since 2010, after Asheesh inspired me <3
- Pyladies co-founder (#pyladies!!)
- Relocating to Berlin (come say Hi!)

Why Scrape?

- So many public APIs and JSON-enabled endpoints (both exposed and not)
- Well-maintained open-source API Libraries
- For python, Selenium is still the best (and really only reliable) bet for anything loaded after the initial page response
- But there are still plenty of sites that don't employ these techniques

What This Talk Will Cover

- LXML vs. BeautifulSoup (with numerous pages)
- Finding Elements within Selenium (which method is fastest)
- Scrapy: How fast can we go?

A Note (Disclaimer)

- There are many other libraries I originally wanted to compare with this, but I found most of them utilized similar functionality or actual dependencies on LXML and BeautifulSoup (html5lib, scrapy)
- I searched widely for “unscrapable” broken pages. I couldn’t find any. If you find one, use BeautifulSoup or html5lib with LXML or cElementTree.
- All of my code for this talk is available at my Github (kjam)

Comparing LXML and BeautifulSoup







- Top libraries for scraping
- Use distinctly different methods for unpacking and parsing HTML
- Both very accurate with the right level of detail (as long as the page is not broken)
- LXML utilizes both xpath as well as cssselect for identifying elements

Methodology

- The methodology I used was to first write accurate scrapers that employed similar techniques of parsing.
- Then I would utilize pstats and cProfile to determine the time and function call. I would then average these over a number of trials (10, 100, 500) to see if there was a distinction.







Case Study: Scraping NHL Scores

ENGLISH | FRANÇAIS | РУССКИЙ | SUOMI | SVENSKA | ČEŠTINA | SLOVENČINA | DEUTSCH


SIGN IN       REGISTER

SCORES | STANDINGS | SCHEDULE | STATS | PLAYERS | TEAMS | NEWS | VIDEO | GAMECENTER LIVE | NHL NETWORK | PLAYOFFS | FANTASY | TICKETS | SHOP


Stay Connected

TRENDING NOW!





Watch the best hockey action happening in the league right now. [WATCH NOW >](#)

 Trending!

- Ryan Miller makes acrobatic save on stomach
- Sam Gagner dazzles in the shootout to win it
- Gustav Nyquist speeds to an electrifying goal
- T.J. Oshie works his magic in the shootout
- Helm completes hat trick with a fancy finish
- Stamkos hits the spin button in the shootout

FS-N **NHL GAMECENTER LIVE WATCH NOW >**



09:22 3rd	1st	2nd	3rd	T
 Minnesota	0	1	0	1
 Winnipeg	0	0	0	0

GOAL SCORERS:
2ND PERIOD:
MIN 01:05 - CHARLIE COYLE (12)

GOALIES:
MIN: I. BRYZGALOV WPG: M. HUTCHINSON

[PREVIEW >](#) [ICE TRACKER >](#) [BOXSCORE >](#) [PHOTOS >](#)



PRIME **NHL GAMECENTER LIVE WATCH LIVE >**

10:00 PM ET	1st	2nd	3rd	T
 Anaheim (50-20-8, 108 PTS)				0
 Vancouver (35-32-11, 81 PTS)				0

[PREVIEW >](#)

FINAL GAMES

NHL GAMECENTER LIVE REPLAY > **PHOTOS >**

FINAL	1st	2nd	3rd	T
 Calgary	0	0	1	1
 New Jersey	0	0	0	0

GOAL SCORERS:
3RD PERIOD:
CGY 00:23 (PPG) - MARK GIORDANO (14)

GOALIES:
CGY: K. RAMO (W) NJD: C. SCHNEIDER (L)

[ICE TRACKER >](#) [BOXSCORE >](#) [RECAP >](#)

“ Three games in a row I thought we played the same, exact way. We played the right way, worked hard, [were] very effective [and] efficient. Everybody contributed and we scored some ugly goals, but we played the right way to get rewarded.

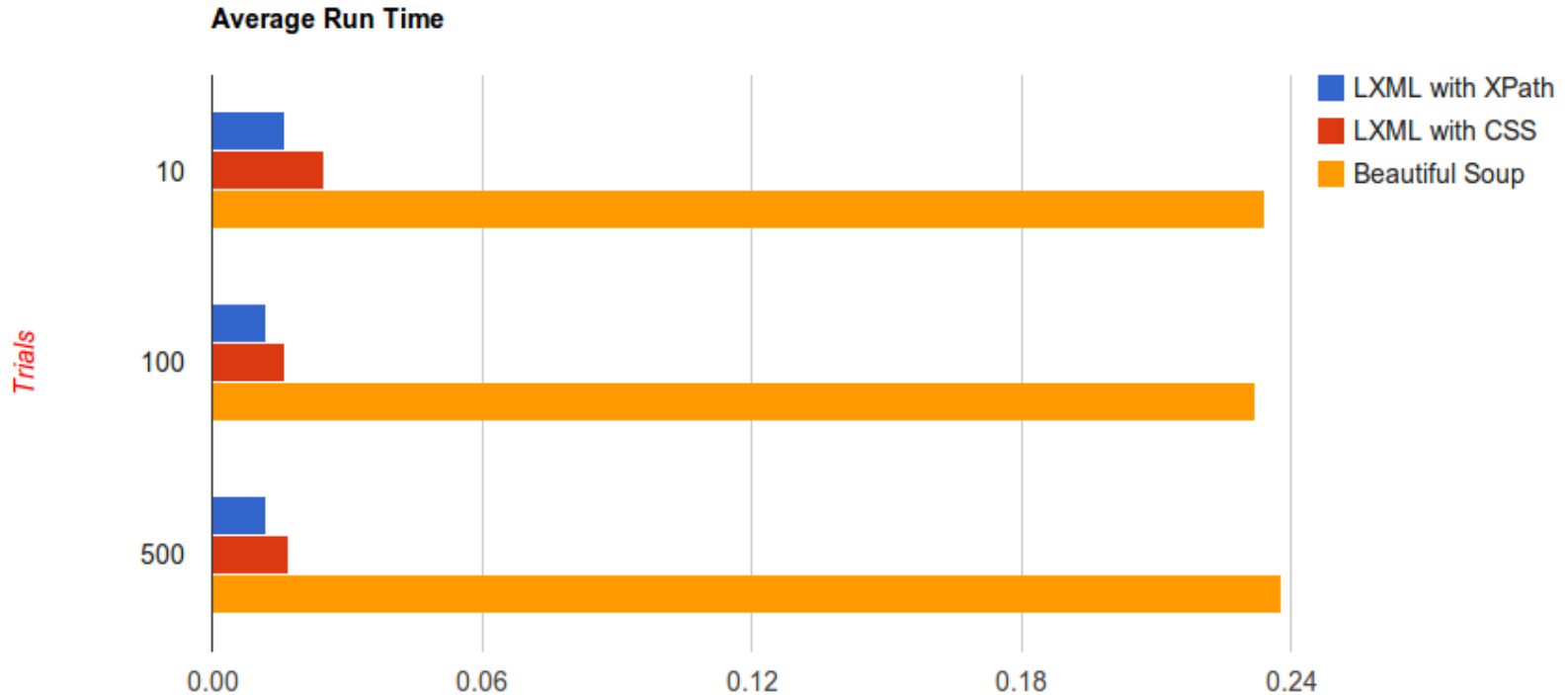
— Blackhawks coach Joel Quenneville after his team's 4-2 win over the St. Louis Blues


```
def run_lxml_xpath():
    all_scores = []
    tree = html.document_fromstring(page)
    scores = tree.xpath('.//div[@id="scoresBody"]')[0]
    for game in scores.xpath('.//div[contains(@class, "gamebox")]'):
        teams = [t.text for t in game.xpath(
            './table//td/a') if t.text]
        totals = [g.text for g in game.xpath(
            './table//td[contains(@class, "total")]')]
        all_scores.append(zip(teams, totals))

def run_lxml_css():
    all_scores = []
    tree = html.document_fromstring(page)
    scores = tree.cssselect('div#scoresBody')[0]
    for game in scores.cssselect('div.gamebox'):
        teams = [t.text_content() for t in game.cssselect('table td.team')]
        totals = [g.text for g in game.cssselect('table td.total')]
        all_scores.append(zip(teams, totals))

def run_beautiful_soup():
    all_scores = []
    tree = BeautifulSoup(page)
    scores = tree.find('div', {'id': 'scoresBody'})
    for game in scores.find_all('div', {'class': 'gamebox'}):
        teams = [t.text for t in game.find_all('td', {'class': 'team'})]
        totals = [tt.text for tt in game.find_all('td', {'class': 'total'})]
        all_scores.append(zip(teams, totals))
```

Case Study: NHL Scores



Case Study: NHL Scores

Library Used	Average Function Calls
LXML with XPath	238
LXML with CSS	2770
Beautiful Soup	280881

Case Study: NHL Scores (Accuracy)

In an accuracy review, all of the scripts accurately found all of the NHL game scores.

Case Study: Scraping Amazon Deals

Today's Deals.

New deals. Every day. Shop our Deal of the Day and more daily deals and limited-time sales.

Never miss another deal

Deal of the Day

\$1.99 Mysteries & Thriller on Kindle


Amazon Digital Services, Inc.

View comments | ★★★★★ (76)

Today only, 50 exciting mysteries and thrillers are only \$1.99 each on Kindle. Kindle books can be read on iPad, iPhone, and Android devices with free Kindle reading apps, as well as Kindle devices

Deal Over

Top Lightning Deal



Omron 7 Series Wrist Blood Pressure Monitor

★★★★★ (2492) | Prime








\$35.99 (59% off)

100% Claimed Deal Over

Today's Lightning Deals

All Available Upcoming Missed Deals

Sort by Category: All

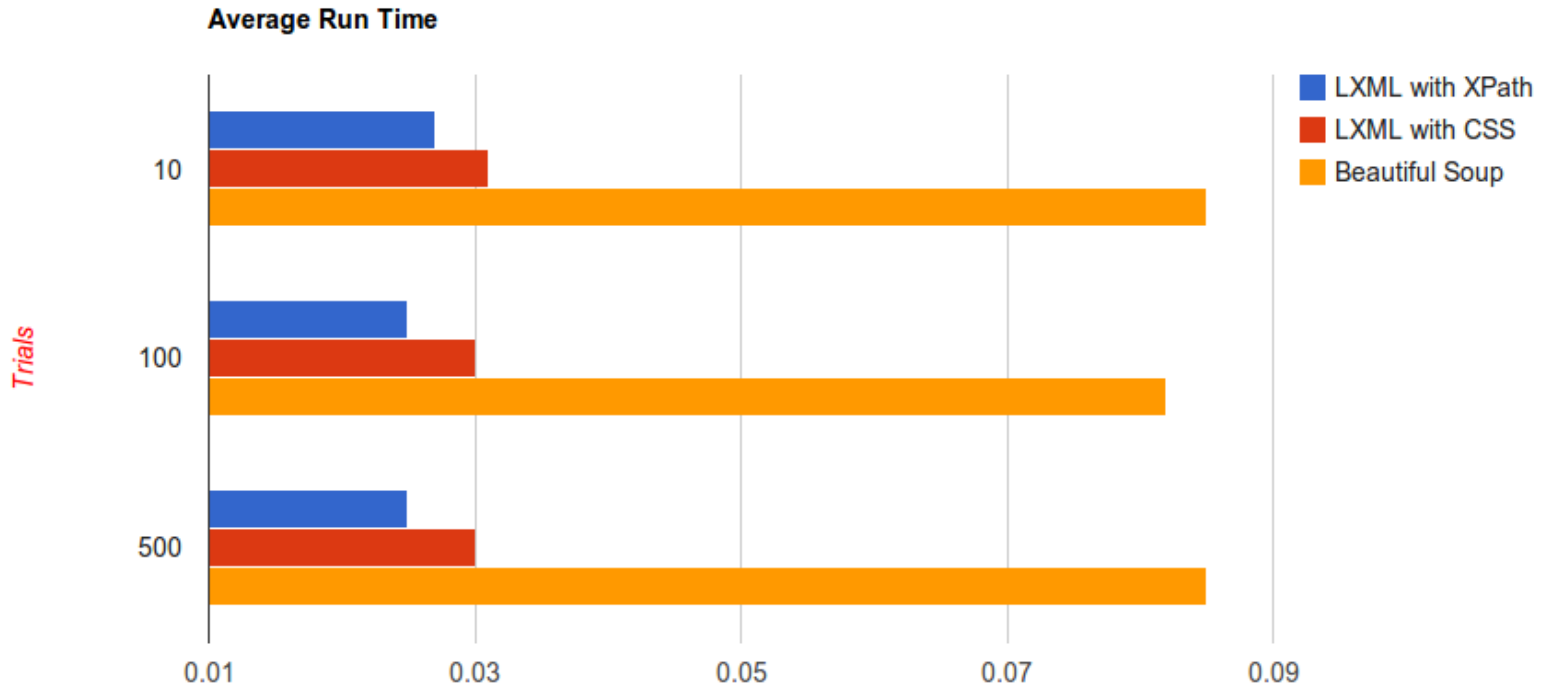
 <p>\$11.19 (28% off)</p> <p>Save on BellyRubs Organic 2 lbs Diatomaceous Earth</p> <p>Oliver's Pet Care</p> <p>72% Claimed Deal Over</p>	 <p>\$16.99 (15% off)</p> <p>Save on West Paw Design Boogey Squeaky Dog Toy</p> <p>Oliver's Pet Care</p> <p>45% Claimed Deal Over</p>	 <p>\$39.99 (69% off)</p> <p>Wahoo Fitness Bike Pack for iPhone</p> <p>9% Claimed Deal Over</p>	 <p>\$199.99 (38% off)</p> <p>Retrospec Bicycles Speck Folding Single-Speed Bicycle</p> <p>100% Claimed Deal Over</p>	 <p>\$95.95 (81% off)</p> <p>Salvatore Extremes Men's 2 Button Striped Charcoal Suit</p> <p>DARYA TRADING INC</p> <p>41% Claimed Deal Over</p>	 <p>\$11.89 (15% off)</p> <p>Save on West Paw Design Kitty Lure Cat Toy</p> <p>Oliver's Pet Care</p> <p>36% Claimed Deal Over</p>	 <p>\$329.95 (59% off)</p> <p>Save over \$450 off List Price on Olympus Pen EPL5</p> <p>Sunset Electronics</p> <p>6% Claimed Deal Over</p>
---	---	---	---	--	---	--

See all deals

Page 1 of 6

Restrictions apply

Case Study: Amazon Deals



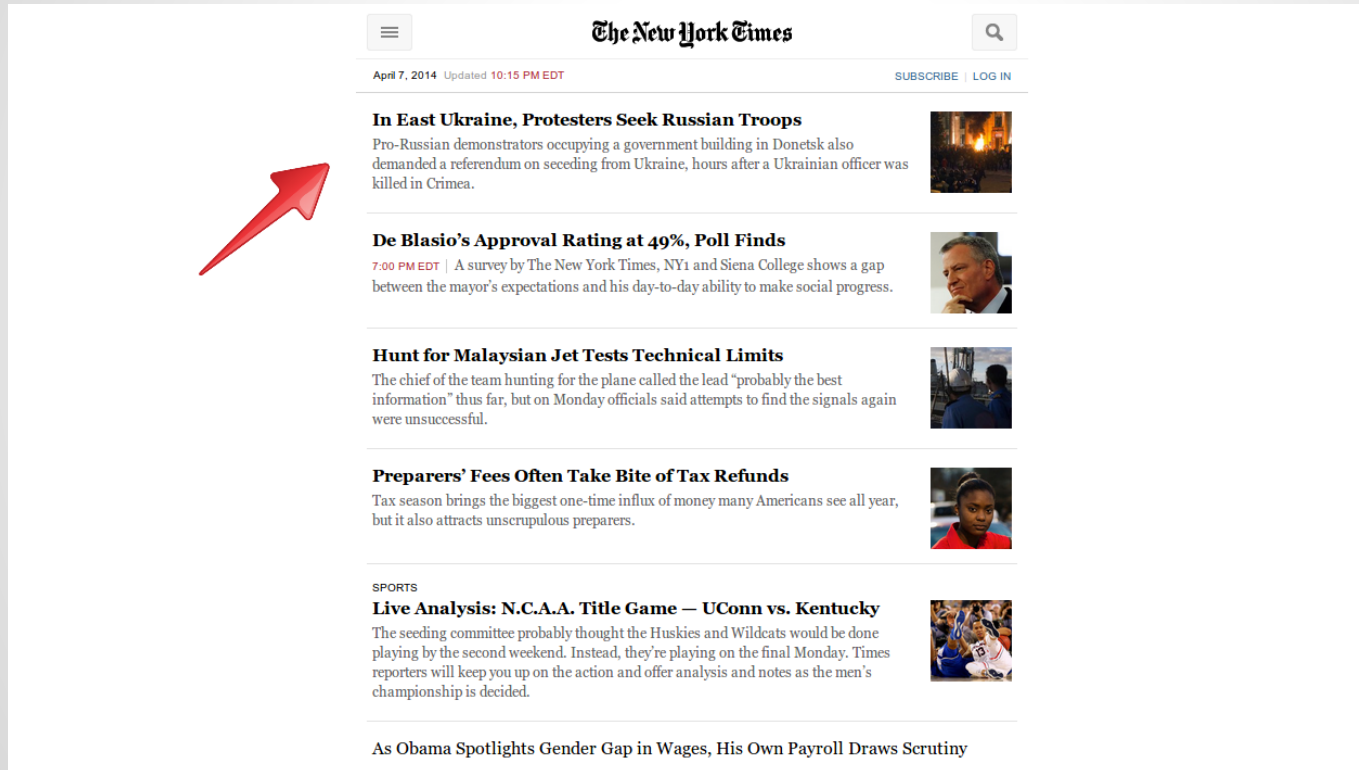
Case Study: Amazon Deals

Library Used	Average Function Calls
LXML with XPath	152
LXML with CSS	1762
Beautiful Soup	86674

Case Study: Amazon Deals

In an accuracy review, BeautifulSoup could not properly parse the more deals section of the page, and therefore I had to modify the BS portion of the scraper to find just the top two deals. I also could not accurately find the price of those deals, so that is omitted for the BS portion of the script.

Case Study: Scraping NYT Mobile



The screenshot shows the mobile interface of The New York Times website. At the top, there is a navigation bar with a hamburger menu icon on the left, the "The New York Times" logo in the center, and a search icon on the right. Below the navigation bar, a date bar indicates "April 7, 2014 Updated 10:15 PM EDT" on the left and "SUBSCRIBE | LOG IN" on the right. The main content area displays a list of news articles. A red arrow points to the first article, "In East Ukraine, Protesters Seek Russian Troops". The article text reads: "Pro-Russian demonstrators occupying a government building in Donetsk also demanded a referendum on seceding from Ukraine, hours after a Ukrainian officer was killed in Crimea." The second article is "De Blasio's Approval Rating at 49%, Poll Finds", with a sub-headline: "7:00 PM EDT | A survey by The New York Times, NY1 and Siena College shows a gap between the mayor's expectations and his day-to-day ability to make social progress." The third article is "Hunt for Malaysian Jet Tests Technical Limits", with a sub-headline: "The chief of the team hunting for the plane called the lead 'probably the best information' thus far, but on Monday officials said attempts to find the signals again were unsuccessful." The fourth article is "Preparers' Fees Often Take Bite of Tax Refunds", with a sub-headline: "Tax season brings the biggest one-time influx of money many Americans see all year, but it also attracts unscrupulous preparers." Below these articles, there is a "SPORTS" section with a sub-headline: "Live Analysis: N.C.A.A. Title Game — UConn vs. Kentucky". The text reads: "The seeding committee probably thought the Huskies and Wildcats would be done playing by the second weekend. Instead, they're playing on the final Monday. Times reporters will keep you up on the action and offer analysis and notes as the men's championship is decided." At the bottom of the page, there is a partial article headline: "As Obama Spotlights Gender Gap in Wages, His Own Payroll Draws Scrutiny".

In East Ukraine, Protesters Seek Russian Troops
Pro-Russian demonstrators occupying a government building in Donetsk also demanded a referendum on seceding from Ukraine, hours after a Ukrainian officer was killed in Crimea.

De Blasio's Approval Rating at 49%, Poll Finds
7:00 PM EDT | A survey by The New York Times, NY1 and Siena College shows a gap between the mayor's expectations and his day-to-day ability to make social progress.

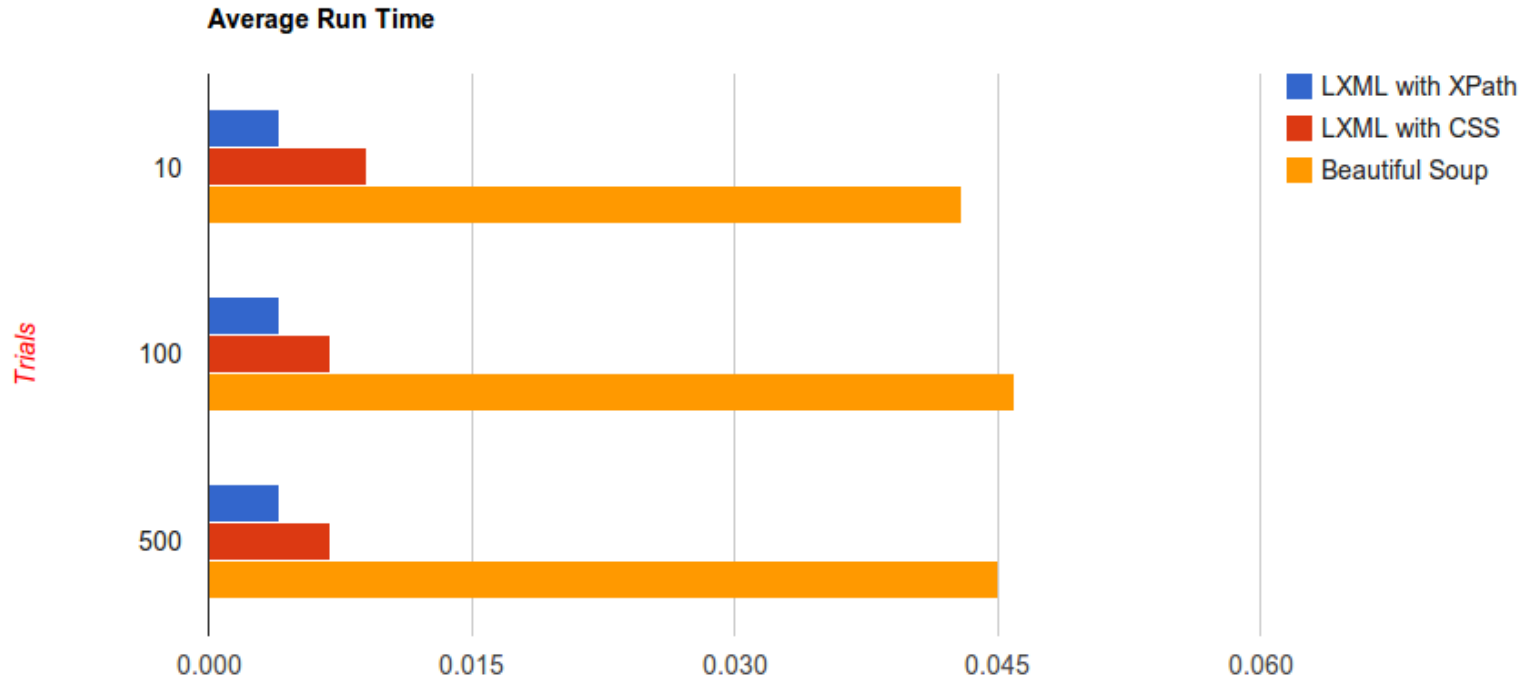
Hunt for Malaysian Jet Tests Technical Limits
The chief of the team hunting for the plane called the lead "probably the best information" thus far, but on Monday officials said attempts to find the signals again were unsuccessful.

Preparers' Fees Often Take Bite of Tax Refunds
Tax season brings the biggest one-time influx of money many Americans see all year, but it also attracts unscrupulous preparers.

SPORTS
Live Analysis: N.C.A.A. Title Game — UConn vs. Kentucky
The seeding committee probably thought the Huskies and Wildcats would be done playing by the second weekend. Instead, they're playing on the final Monday. Times reporters will keep you up on the action and offer analysis and notes as the men's championship is decided.

As Obama Spotlights Gender Gap in Wages, His Own Payroll Draws Scrutiny

Case Study: NYT Mobile



Case Study: NYT Mobile

Library Used	Average Function Calls
LXML with XPath	345
LXML with CSS	1799
Beautiful Soup	47733

Case Study: NYT Mobile

In an accuracy review, all of the scripts found 17 articles on the page, including an empty set at the bottom.

LXML with XPath!

- Clear winner!
- But at the end of the day, not by much. :)

Let's investigate Selenium

- Best library for page interactions and after DOM load elements
- There are *many* ways to find elements on a page. Which is the fastest?
- I'm going to compare tag_name, class_name (css) and XPath.

Selenium: Comparing Element Find

The image shows a screenshot of the Yahoo! homepage. At the top, there is a navigation bar with links for Home, Mail, News, Sports, Finance, Weather, Games, Groups, Answers, Screen, Flickr, Mobile, and More. Below this is the Yahoo! logo and a search bar with a blue 'Search' button. A red arrow points to the search bar. On the left side, there is a vertical menu with icons for Mail, Autos, News, Sports, Finance, Weather, Games, Homes, Food, Tech, Answers, Screen, Flickr, Jobs, Shopping, Travel, and Dating. The main content area features a large image of a Boeing X-37B space plane with the headline "Secret plane has been in space for nearly 500 days". Below this is a carousel of smaller images with headlines: "Air Force's secret plane", "Bizarre sea creatures", "Royal family Down Under", "Bieber given \$1.9M Bugatti", and "Chill's stir controversy". To the right, there is a "Trending Now" section with a list of 10 items. A red arrow points to the 10th item, "Windows XP". Below that is the "NCAA Tourney" section, showing a bracket for the National Final between Kentucky and Connecticut. At the bottom, there is a "Today's Headlines" section with five items, including "Sub hunting for source of 'pings' in plane search".

Home Mail News Sports Finance Weather Games Groups Answers Screen Flickr Mobile More

YAHOO!

Search

My Yahoo Sign In Mail

Mail
Autos
News
Sports
Finance
Weather
Games
Homes
Food
Tech
Answers
Screen
Flickr
Jobs
Shopping
Travel
Dating

More Yahoo Sites >

Get the Yahoo mobile app

Secret plane has been in space for nearly 500 days
The Air Force still isn't saying just what exactly the X-37B is up to — or when it's coming back to Earth. [Some hypotheses >](#) 1 - 5 of 85

Air Force's secret plane
Bizarre sea creatures
Royal family Down Under
Bieber given \$1.9M Bugatti
Chill's stir controversy

2014 NCAA MEN'S TOURNAMENT
Injured Kentucky star's shirt raises eyebrows
Willie Cauley-Stein appears to have a little Dennis Rodman in him.
[Live game updates](#) [Photos](#) [Complete tourney coverage](#)

All Stories News Local Entertainment Sports More >

Kevin Federline Has a Six Pack! Wife Victoria Prince Gives Birth to a Girl
Kevin Federline may not have much of a six pack left around his waist these days, but he has one in the kid department.
Yahoo Celebrity

Trending Now [Watch the show >](#)

- 1 Jeffrey Dahmer
- 2 Jamie Lynn Spears
- 3 Malaysia Airlines flig...
- 4 WWE WrestleMania
- 5 Santa Barbara riot
- 6 Amanda Knox
- 7 Taraji P. Henson
- 8 Duck Commander 500
- 9 Doctor Zhivago
- 10 Windows XP

NCAA Tourney National Final

(6) Kentucky 34 - 35 (7) Connecticut
19:16 2nd

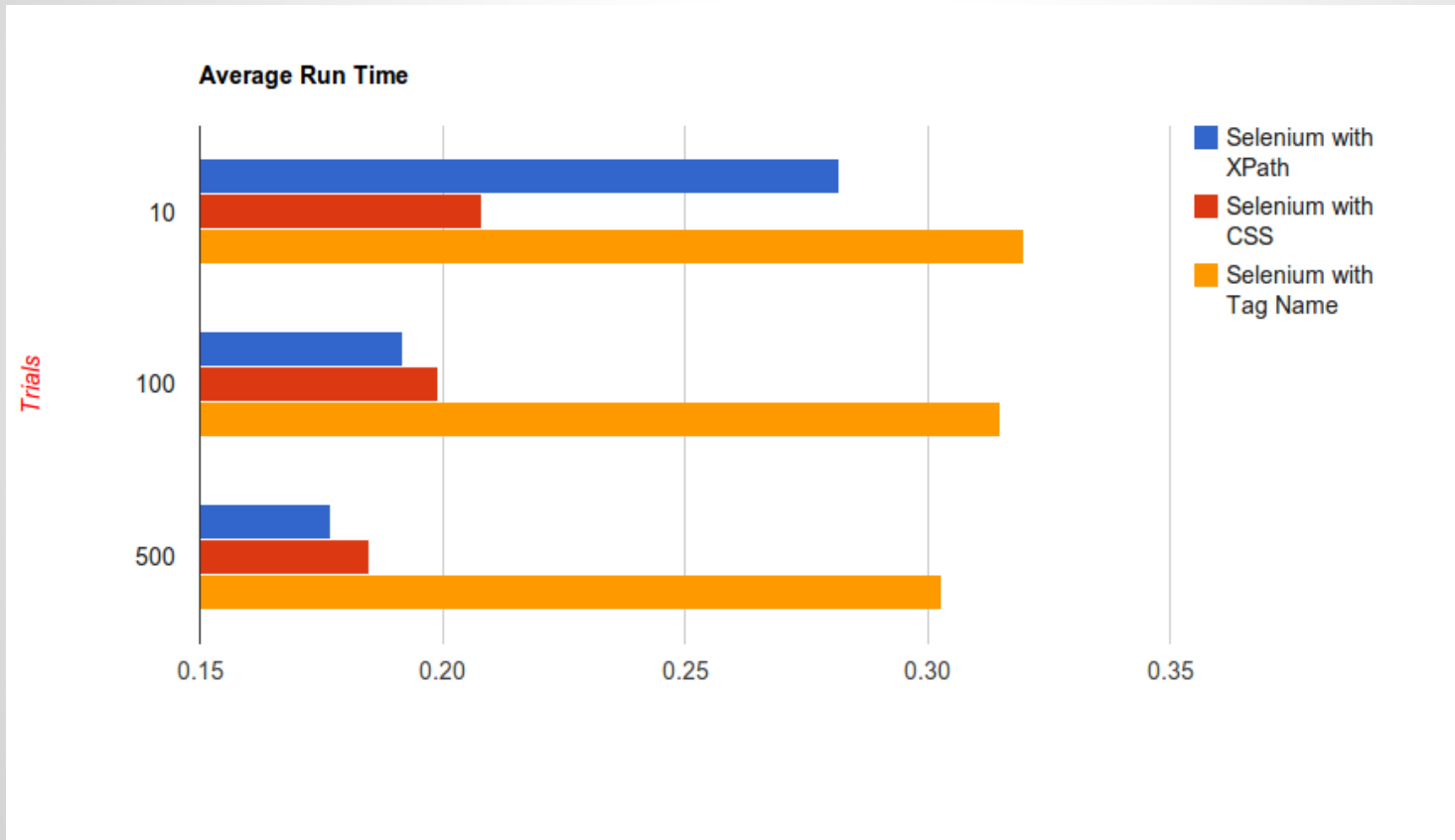
How busted is your bracket?
[Complete tournament coverage](#)

Today's Headlines

- 1 Sub hunting for source of 'pings' in plane search
14 minutes ago
- 2 Army: Ft. Hood shooting suspect had requested leave
- 3 Citigroup to pay \$1.125B in mortgage settlement
- 4 Senate OKs jobless bill; House prospects slimmer
- 5 Water woes reportedly endanger North Korea reactor

Sports

Selenium: A Speed Comparison



Selenium: Function Calls

Library Used	Average Function Calls
Find with XPath	11880
Find with CSS	2980
Find with Tag Name	12881

Tag Name: Clear Loser

- CSS and XPath are both great
- Tag is clearly slower and with more calls
- Similarly to web scraping, it's not *that* huge of a difference; so always use what works best for your script and something you find comfortable and readable.

Let's investigate Scrapy

- Utilizes lxml XPath for finding elements (or items)
- Utilizes Twisted for asynchronous crawling
- Best library by far in terms of crawling or spidering the web
- With our speed knowledge, obvious choice for parsing a series of pages with speed
- **How fast can we go?**

Scrapy: LXML Speed with Twisted

- Test: Query Google with pagination for search results
- Find items that have title, blurb, link. I didn't worry about writing it somewhere, so that would have added time, but I did create objects
- I googled "python" (because why not?)

Scrapy Stats

```
2014-04-11 09:15:57-0400 [google_results] INFO: Dumping Scrapy stats:
  {'downloader/request_bytes': 13093,
   'downloader/request_count': 36,
   'downloader/request_method_count/GET': 36,
   'downloader/response_bytes': 1467915,
   'downloader/response_count': 36,
   'downloader/response_status_count/200': 36,
   'finish_reason': 'finished',
   'finish_time': datetime.datetime(2014, 4, 11, 13, 15, 57, 491256),
   'item_scraped_count': 306,
   'log_count/DEBUG': 344,
   'log_count/ERROR': 5,
   'log_count/INFO': 7,
   'response_received_count': 36,
   'scheduler/dequeued': 36,
   'scheduler/dequeued/memory': 36,
   'scheduler/enqueued': 36,
   'scheduler/enqueued/memory': 36,
   'spider_exceptions/TypeError': 5,
   'start_time': datetime.datetime(2014, 4, 11, 13, 15, 54, 466743)}
2014-04-11 09:15:57-0400 [google_results] INFO: Spider closed (finished)
Fri Apr 11 09:15:57 2014      stats
```

702840 function calls (682062 primitive calls) in 6.044 seconds

Scrapy: Scraping Google

- Spider was averaging ~ 100 results / second!
- Google now hates me
- Scrapy has a lot of different tools to get around things like Google captcha block, but I didn't invest the time into playing with it to get it working 100% of the time, but please feel free to fork and do so! :)

In Conclusion

- LXML using XPath is the clear winner when it comes to speed.
- Readability and accuracy (both in your code and in the content you scrape) is pretty key as well. Your use might vary from these tests but keep it in mind.
- If XPath is too confusing or limiting, cssselect appears to be a close second in speed.

Any Questions?

- Ask now!
- Ask later:
 - @kjam on twitter
 - /msg kjam on Freenode
- Thanks! :D