Web Scraping

Scratch your itch.

Hi. I'm Greg Reda.

gregreda.com

@gjreda

Problem

Some websites present information poorly.

Or you have a problem/question and need data.

You searched for:
2011 ☑
Introduction ☑

START OVER

Narrow Search

Year

2011

Section

Introduction

Show only

- Readers' Poll
- Critics' Picks
- Image

Category

- Best of Chicago 2011 Winners
- Goods & Services
- Music & Nightlife
- Sports & Recreation
- Food & Drink
- Arts & Culture
- City Life

Best of Chicago 2011

Goods & Services »

BEST OF CHICAGO 2011 WINNERS

If there ever was a time and place for offering a nod to the old guard while plowing ahead into an uncertain and exciting future, the time is now and the place is Chicago.

Readers' Poll Winners

Best restaurant that's been around forever and is still worth the trip

Best fancy restaurant in Chicago

Best bang for your buck

Best chef

Best up-and-coming chef

Best food blog

Best ampersand restaurant

Best restaurant name

Best new food trend

Best cocktail list

Best mixologist

Best wine list

Best sommelier

Best brewpub

Best local brew

Best wine shop

Best liquor store

Best BYOB

Best alfresco dining

Best late night

Best for kids

Best waitstaff

Best-looking waitstaff

Best food festival

Best food truck

Best gourmet market

Best local grocer

Best local food product

Best farmers market

Best butcher shop

Best cheesemonger

Reader Critics' Picks

Best Italian steak house where my dad felt at home in the 60s and I do now

Best case of nostalgia, bordering on time travel

Best restaurant empire founded the same year as the Reader

Best restaurant

Best bargain Michelin chef

Best chef downshift, animal division

Best chef downshift, vegetable division

Best venerable restaurant alongside the el

Best new food truck

Best buffet

Best game day

Best dairy product to camp out in front of the cheese shop for

Best use of alcohol at breakfast

Best university coffeehouse

Best bakery you've never heard Of

Best place to see bakers at work

Best place for ambience and egg sandwiches

Best bagel

Best tubular collaboration

Best sausage

Best place in Chicago to sample salumi from Mario Batali's papa

Best broccoli and shells con patio

Best fancy-pants pizza special

Best Polish-Mexican-American $1.50 taco

Best tater tots

Best spinach pie

Best som tam

Best soundtracked strawberry shake

Every category looked like this.

And every winner looked like this.

You searched for:

2011 ☑

Food & Drink ☑

START OVER

Narrow Search

Year

2011

Section

Food & Drink

Show only

- Readers' Poll
- Critics' Picks
- Image

Category

- Best restaurant that's been around forever and is still worth the trip
- Best fancy restaurant in Chicago
- Best bang for your buck
- Best chef

Best of Chicago 2011 » Food & Drink

« Best fancy restaurant in Chicago

Best chef »

BEST BANG FOR YOUR BUCK

TIE

BIG STAR

1531 N. Damen

773-680-7740

bigstarchicago.com

SULTAN'S MARKET

2057 W. North and other locations

773-235-3072

chicagofalafel.com

1. **BLT**
   Old Oak Tap
   Read more

2. **Fried Bologna**
   Au Cheval
   Read more

3. **Woodland Mushroom**
   Xoco
   Read more

4. **Roast Beef**
   Al's Deli
   Read more

5. **PB&L**
   Publican Quality Meats
   Read more

6. **Belgian Chicken Curry Salad**
   Hendrickx Belgian Bread Crafter
   Read more

7. **Lobster Roll**
   Acadia
   Read more

8. **Smoked Salmon Salad**
   Birchwood Kitchen
   Read more

- The 606 Shows How to Design a Park in the 21st Century (and Beyond)
- A Chapel Conversion in Logan Square Goes Under Contract in Two Days
- Sink|Swim's Cocktails Are Worth the Trip
- See Chicago from Above in These Stunning Photos
- How to Ge Around the 606
- Nick Offe an Will Sing at the Green Mill on Sat da
- The Yupp Tu s S

rank: 35
sandwich: B.A.D.
restaurant: Floriole Cafe & Bakery
description: The silky smooth almond-date spread makes this a standout. No, wait, it's that perfect bacon. Actually, it could be the wonderfully grainy housemade cornmeal bread. Whatever. This sandwich doesn't live up to its name.
price: $9
address: 1220 W. Webster Ave.
city: Chicago
phone: 773-883-1313
website: floriole.com
full_address: 1220 W. Webster Ave., Chicago

Better.

Solution

Scrape it.

But be nice.

2012 BEST SANDWICHES

1. Old Oak Tap BLT

"Truly inspired."

PUBLISHED OCT. 9, 2012

PHOTO: ANNA KNOTT; FOOD STYLIST: LISA KUEHL

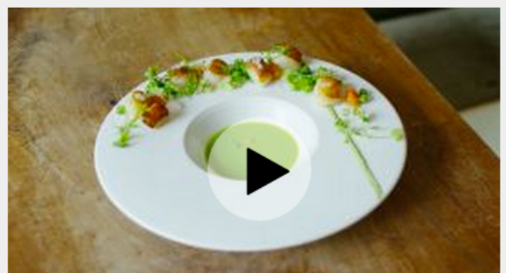
The B is applewood smoked—nice and snappy. The L is arugula—fresh and peppery. The T is a fried green slice—jacketed in cornmeal and greaseless. Slathered with pimiento cheese, the grilled ciabatta somehow stays crisp, providing three distinct layers of crunch. Truly inspired.

$10. 2109 W. Chicago Ave., 773-772-0406, theoldoaktap.com

← PREVIOUS
50. The Gatsby
Phoebe's Bakery

NEXT →
2. Fried Bologna
Au Cheval

This article appears in the November 2012 issue of *Chicago* magazine. Subscribe to *Chicago* magazine.



Check the page source.

The B is applewood smoked—nice and snappy. The L is arugula—fresh and peppery. The T is a fried green slice—jacketed in cornmeal and greaseless. Slathered with pimiento cheese, the grilled ciabatta somehow stays crisp, providing three distinct layers of crunch. Truly inspired.

$10. 2109 W. Chicago Ave., 773-772-0406, theoldoaktap.com

← PREVIOUS
NEXT →

How?

Python + BeautifulSoup

HTML + CSS

```
<body>
  <h1 class="headline">1. Old Oak Tap BLT</h1>
  <p>The B is applewood smoked&#8212;nice and snappy. The L is
arugula&#8212;fresh and peppery. The T is a fried green
slice&#8212;jacketed in cornmeal and greaseless. Slathered with
pimiento cheese, the grilled ciabatta somehow stays crisp,
providing three distinct layers of crunch. Truly inspired.</p>
  <p class="addy">
    <em>$10. 2109 W. Chicago Ave., 773-772-0406, <a
href="http://www.theoldoaktap.com/">theoldoaktap.com</a></em>
  </p>
</body>
```

Python + bs4

```
from bs4 import BeautifulSoup
import requests

url = ('http://www.chicagomag.com/Chicago-Magazine/November-2012/'
        'Best-Sandwiches-in-Chicago-Old-Oak-Tap-BLT/')
r = requests.get(url)
soup = BeautifulSoup(r.text)

sandwich = soup.find('h1', class_='headline').get_text()
desc = soup.find_all('p')[0].get_text()

# ugly code FTW - actually what I wrote
addy = soup.find('p', class_='addy').em.get_text()
foo = addy.split(',')[0].strip()
price = foo.partition(" ")[0].strip()
address = foo.partition(" ")[2].strip()
```

What about Javascript?

You're lucky!

Check your browser's network tab.

PTS 20 REB 5.8 AST 3.3 PIE 13.8

COMPARE PLAYER

PROFILE | STATS PROFILE | STATS | CAREER | GAME LOGS | TRACKING

2014-15 PLAYOFFS SHOT LOGS

Page 1 of 5 | 213 Rows

Game	Loc	W/L	Final Margin	Shot #	Period	Game Clock	Shot Clock	Dribbles	Touch Time	Shot Dist	PTS Type	Shot Result	Closest Defender	Defender Dist
MAY 14, 2015 - CHI vs. CLE	H	L	-21	1	1	8:44	24	0	0	2.6	2	made	Mozgov, Timofey	2.6

Elements | Network | Sources | Timeline | Profiles | Resources | Audits | Console

3 2

View: Options: Preserve log Disable cache

Filter

All | XHR | Script | Style | Images | Media | Fonts | Documents | WebSockets | Other Hide data URLs

Name

summary.html
trackingLogsShots.html
commonplayerinfo?Leag...
playerdashptshotlog?Dat...
YLs2gGY4Xkc.js

5 / 96 requests | 9.1 KB / 1...

× Headers | Preview | Response | Cookies | Timing

General

Remote Address: 23.0.160.40:80

Request URL: http://stats.nba.com/stats/playerdashptshotlog?DateFrom=&DateTo=&GameSegment=&LastNGames=0&LeagueID=00&Location=&Month=0&OpponentTeamID=0&Outcome=&Period=0&PlayerID=202710&Season=2014-15&SeasonSegment=&SeasonType=Playoffs&TeamID=0&VsConference=&VsDivision=

Request Method: GET

Status Code: 200 OK

Response Headers view source

Cache-Control: no-cache, no-store, must-revalidate

Connection: keep-alive

Content-Encoding: gzip

Content-Length: 4879

Python + Requests

```
import requests

url = 'http://stats.nba.com/stats/playerdashptshotlog'
headers = { ... }
params = {'PlayerID': 202710, 'Season': '2014-15', ... }
r = requests.get(url, headers=headers, params=params)

if r.ok:
    do_something(r.json())
```

Remember

Data science is about *answering questions* and *solving problems*.

Scratch your own itch.

Tools

Python + BeautifulSoup (or Scrapy)

R + rvest

Ruby + Nokogiri

...

Resources

Web Scraping 101: Getting Started

Web Scraping 201: Finding the API

Web Scraping in R

Steal These Projects

BeerMenus + Beer Ratings Mashup

NFL Pass Interference over time

Best late game basketball coaches

Thank you.