

Web Scraping

CS59I - Data Science with Python

Slides adapted from March 2015
Web Scraping lecture by Davide Proserpio

Today's material

cs-people.bu.edu/lspinel/cs/cs591.zip

What is web scraping?

Is a technique used to **extract data** from websites
(also called web harvesting or data extraction)

Best prices for: 1 room v 2 guests v

03/20/2015



03/30/2015

Book on
tripadvisor**\$429*** >

\$62 taxes & fees

Expedia

\$416* >

\$60 taxes & fees

travelocity

\$416* >

\$60 taxes & fees

Booking.com

\$429*

amotxtravel.com

\$429*

Orbitz.com

\$383*

10 more sites v

*Disclaimer

Luxury

Pets Allowed

Fenway / Kenmore

Traveler
photosProfessional
photosBrowse
nearby

2,916 reviews from our community

Write a Review

Traveler rating

Excellent		2,360
Very good		434
Average		81
Poor		39
Terrible		12

See reviews for

	Families	902
	Couples	847
	Solo	146
	Business	682

Rating summary

Sleep Quality	
Location	
Rooms	
Service	
Value	
Cleanliness	

Related hotels...

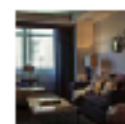
**XV Beacon**

793 Reviews

\$480 and up* v**Boston Harbor Hotel**

1,263 Reviews

Lowest price at Expedia, 0 sites checked

\$358 and up* v**Nine Zero Hotel - a Kimpton Hotel**

1,805 Reviews

Lowest price at Expedia, 6 sites checked

Best prices for: 1 room v 2 guests v

03/20/2015



03/30/2015

Book on
tripadvisor**\$429***

\$62 taxes & fees



Expedia

\$416*

\$60 taxes & fees



travelocity

\$416*

\$60 taxes & fees

Booking.com
Orbitz.com**\$429***
\$283*

amextravel.com

\$429*

10 more sites v

*Disclaimer

Luxury

Kenmore

Traveler
photosProfessional
photosBrowse
nearby

Back

Forward

Reload

Save As...

Print...

Translate to English

View Page Source

View Page Info

AdBlock ▶

JSONView ▶

Inspect Element

2,916 reviews from our community

Traveler rating

Excellent		2,350
Very good		434
Average		81
Poor		39
Terrible		12

See reviews for

	Families	902
	Couples	847
	Solo	145
	Business	582

Rating summary

Sleep Quality	
Location	
Rooms	
Service	
Value	
Cleanliness	

Related hotels...

**XV Beacon**

753 Reviews

\$480 and up***Boston Harbor Hotel**

1,233 Reviews

Lowest price at Expedia, 8 sites checked

\$358 and up***Nine Zero Hotel - a Kimpton Hotel**

1,005 Reviews

Lowest price at Expedia, 8 sites checked


```

1000 </li>
1001 <li class="tabItem" href="#><a onclick="ta.setEvtCookie('TopNav', 'click', 'Flights', 0, this.href);setPID(3820)" href="/Flights-g60745-Boston_Massachusetts-
Cheap Discount Airfare.html" class="tabLink pid3820">
1002 Flights
1003 </li>
1004 <li class="tabItem" href="#><a onclick="ta.setEvtCookie('TopNav', 'click', 'VacationRentals', 0, this.href)" href="/VacationRentals-g60745-Reviews-
Boston_Massachusetts-Vacation_Rentals.html" class="tabLink pid2795">
1005 Vacation Rentals
1006 </li>
1007 </li>
1008 <li class="tabItem dropDown jsNavMenu href="#>
1009 <a onclick="ta.util.cookie.setPIDCookie(4467); ta.setEvtCookie('TopNav', 'click', 'Restaurants', 0, this.href)"
recommended="ta.common.header.addClearParam(this);" href="/Restaurants-g60745-Boston_Massachusetts.html" class="tabLink subLink ">Open
class="arrow_text">Restaurants</a><div class="arrow_dropdown_wht sprite-arrow_dropdown_wht_refresh" src="http://a2.tacdn.com/img2/x.gif" alt="" width="3"
height="7"></div>
1010 <ul class="subNav">
1011 <li class="subItem">
1012 <a class="subLink" href="/Restaurants-g60745-Boston_Massachusetts.html" onMouseover="ta.common.header.addClearParam(this);">All Boston Restaurants</a> </li>
1013 <li class="subItem">
1014 <a class="subLink" href="/RestaurantsNear-g60745-d218705-Hotel_Commonwealth-Boston_Massachusetts.html">Restaurants near Hotel Commonwealth</a>
1015 </li>
1016 </ul>
1017 <li class="tabItem dropDown jsNavMenu href="#>
1018 <a onclick="ta.util.cookie.setPIDCookie(4467); ta.setEvtCookie('TopNav', 'click', 'ThingsToDo', 0, this.href)" href="/Attractions-g60745-Activities-
Boston_Massachusetts.html" class="tabLink subLink ">Open class="arrow_text">Things to Do</a><div class="arrow_dropdown_wht sprite-
arrow_dropdown_wht_refresh" src="http://a2.tacdn.com/img2/x.gif" alt="" width="3" height="7"></div>
1019 <ul class="subNav">
1020 <li class="subItem">
1021 <a class="subLink" href="/Attractions-g60745-Activities-Boston_Massachusetts.html" onMouseover="ta.common.header.addClearParam(this);">All things to do in
Boston</a> </li>
1022 <li class="subItem">
1023 <a class="subLink" href="/AttractionsNear-g60745-d218705-Hotel_Commonwealth-Boston_Massachusetts.html">Things to do near Hotel Commonwealth</a>
1024 </li>
1025 </ul>
1026 </li>
1027 <li class="tabItem" href="#><a onclick="ta.setEvtCookie('TopNav', 'click', 'TravelersChoice', 0, this.href)" href="/TravelersChoice" class="tabLink pid5087">
1028 Best of 2015
1029 </li>
1030 </li>
1031 <li class="tabItem dropDown jsNavMenu href="#>
1032 <a href="#" class="tabLink subLink"><div class="arrow_text">More</div><div class="arrow_dropdown_wht sprite-arrow_dropdown_wht_refresh"
src="http://a2.tacdn.com/img2/x.gif" alt="" width="3" height="7"></div>
1033 <ul class="subNav">
1034 <li class="subItem">
1035 <a href="/Travel_Guide-g60745-Boston_Massachusetts.html" onclick="ta.setEvtCookie('TopNav', 'click', 'TravelGuides', 0, this.href)" class="subLink
pid1615">Travel Guides
1036 </li>
1037 <li class="subItem">
1038 <a href="/ShowForum-g60745-148-Boston_Massachusetts.html" onclick="ta.setEvtCookie('TopNav', 'click', 'TravelForum', 0, this.href)" class="subLink
pid14621">Travel Forum
1039 </li>
1040 </ul>
1041 <li class="subItem">
1042 <a href="/apps" onclick="ta.setEvtCookie('TopNav', 'click', 'Apps', 0, this.href)" class="subLink pid18876">Apps
1043 </li>
1044 </li>
1045 <li class="subItem">
1046 <a href="/GreenLeaders" onclick="ta.setEvtCookie('TopNav', 'click', 'GreenLeaders', 0, this.href)" class="subLink pid34563">GreenLeaders
1047 </li>
1048 </li>
1049 <li class="subItem">
1050 <a href="/GreenLeaders" onclick="ta.setEvtCookie('TopNav', 'click', 'GreenLeaders', 0, this.href)" class="subLink pid34563">GreenLeaders
1051 </li>
1052 </li>

```

Web scraping I

1. Retrieve webpages:

- low-level HTTP
 - wget, curl: command line tools and library for transferring data with URL syntax
- Fully-fledged web browsers
 - Selenium (web browser automation)

Web scraping II

2. Parse and extract information from the html

- Html parsers (e.g. Python BeautifulSoup)

```
<a class="subLink" href="/Hotels-g60745-zff4-Boston_Massachusetts-Hotels.html">Boston Family Hotels</a>
```

- JSON parsers (e.g. Python json)

```
{  
  url_type: "geo",  
  name: "Boston Tourism",  
  type: "GEO",  
  url: "/Tourism-g60745-Boston_Massachusetts-Vacations.html"  
}
```


Web scraping II

3. Store the data: MySQL, SQL, etc



Retrieve Webpages I

Embed command line tools (wget or curl) in the python code:

```
import os
# define output file name
>>out_file = "test.html"
# url page
>>url = www.tripadvisor.com
# crate the cURL command (as a string)
>>cmd = "curl -L -m 20 '%s' > %s" % (url,
out_file)
# use os libray to retrieve tha page
>>os.system(cmd)
```

Retrieve Webpages II

Use a python network-access libraries:

- Urllib, Requests, pycurl, etc.

```
# Retrieve the page
```

```
>>response = requests.get(url)
```

```
# Get the html
```

```
>>html = response.text.encode('utf-8')
```

Parse and extract data from HTML: Beautifulsoup

```
>>soup = BeautifulSoup(html_file)
>>soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

>>soup.find('a', id="link3")
# <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>

>>for link in soup.find_all('a'):
    print(link.get('href'))
# http://example.com/elsie
# http://example.com/lacie
# http://example.com/tillie
```

Parse and extract data form JSON: Python json

JSON file json-test.txt:

```
{  
url_type: "geo",  
name: "Boston Tourism",  
type: "GEO",  
url: "/Tourism-g60745-Boston_Massachusetts-Vacations.html"  
}
```

```
# Read file  
>>json_file = open("json-text.txt", 'r')  
# Load it using json library  
>>js_obj = json.loads(json_file.read())  
# Retrieve name value  
>>Name = js_obj['name']
```

Common practice I

When [collecting](#) the data:

- “*Inspect element*”: helpful to get the format of the HTTP request + headers.
- Do NOT make HTTP requests too fast
 - you can be blocked
 - ... but you can use web proxies (e.g. <https://www.hidemyass.com/>)
- [Mimic](#) the “*real*” http request as much as possible
 - Set HTTP headers: User-agent, Host, Referer, etc.
- Use Website [API](#) (if exists) (e.g. [Airbnb API](#))

Common practice II

When parsing the data:

- “*Inspect element*”: very helpful when writing the parser
- Note that sometimes code downloaded != code in your browser

(always save your html files so that you can explore them)

Disclaimer: Legal issues

Web scraping may be **against** the terms of use of some websites!



PROHIBITED ACTIVITIES

The content and information on this Website (including, but not limited to, messages, data, information, text, music, sound, photos, graphics, video, maps, icons, software, code or other material), as well as the infrastructure used to provide such content and information, is proprietary to us. You agree not to otherwise modify, copy, distribute, transmit, display, perform, reproduce, publish, license, create derivative works from, transfer, or sell or re-sell any information, software, products, or services obtained from or through this Website. **Additionally, you agree not to:**




- (i) use this Website or its contents for any commercial purpose
- (ii) **access, monitor or copy any content or information of this Website using any robot, spider, scraper or other automated means or any manual process for any purpose without our express written permission;**

Today's example

1. Scrape list of Boston hotels from TripAdvisor
2. Extract (and print) the following information:
 - Hotel Name
 - Average rating
 - Number of reviews

Step 1

Obtain the TripAdvisor Boston page


 know better  book better  go better

[Hotels](#) [Flights](#) [Vacation Rentals](#) [Restaurants](#) [Things to do](#) [Best of 2016](#) [More](#)



Find: Hotels, Restaurants, Things to Do **Near:** Enter a destination **Search**

Special Offer:
Book a hotel on TripAdvisor + **Win** a \$1,500 getaway!

HOTELS **VACATION RENTALS** **THINGS TO DO** **RESTAURANTS** **FLIGHTS**

 Enter destination or hotel name

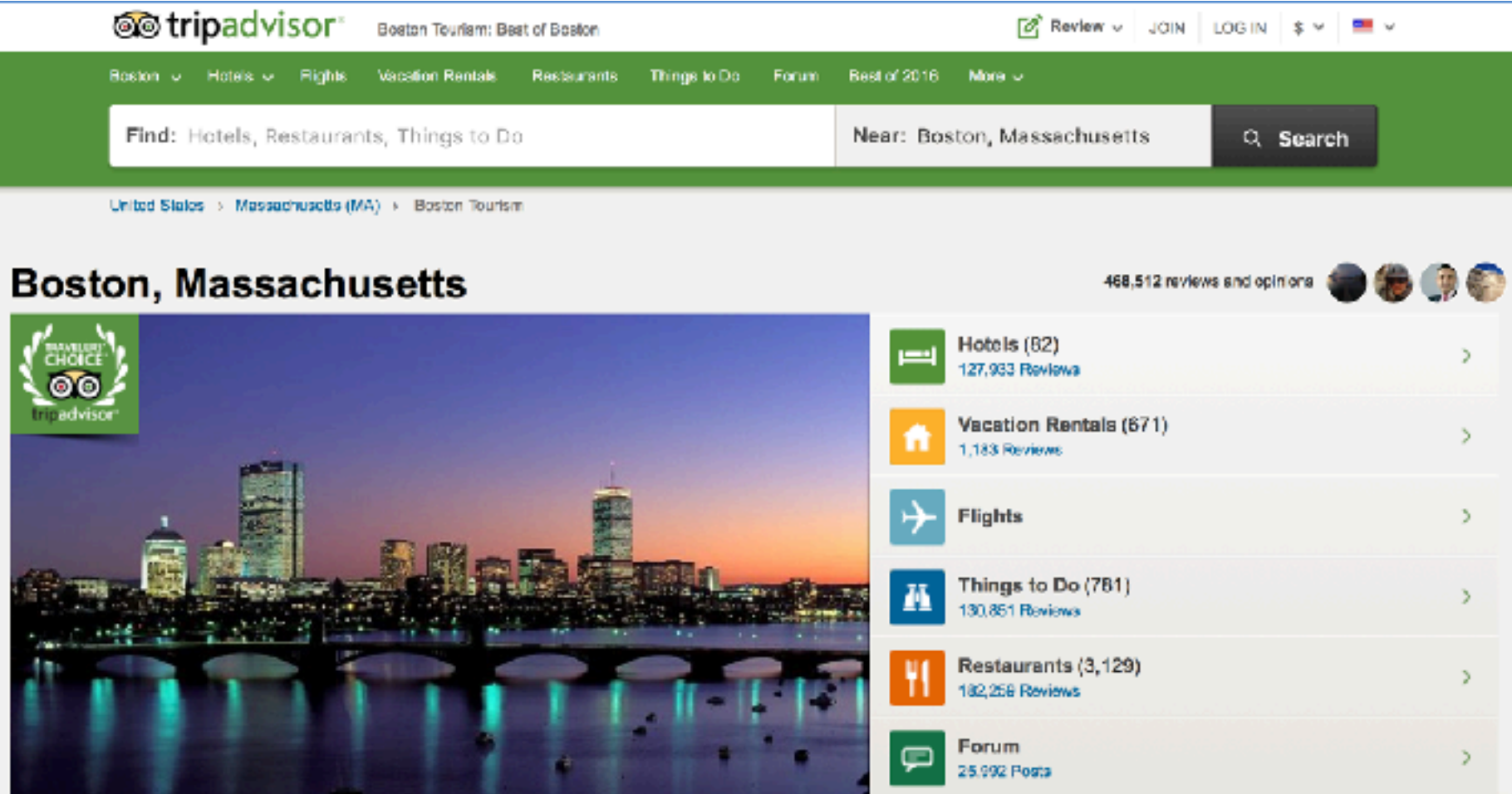
Find Hotels

 *"Heaven on earth"*
 Review by 3Guapas
See all 40,061 reviews of Provinciales

Step 1

Obtain the TripAdvisor Boston page

Get: <https://www.tripadvisor.com/Boston>



The screenshot shows the TripAdvisor website for Boston. The header includes the TripAdvisor logo, the text "Boston Tourism: Best of Boston", and links for "Review", "JOIN", "LOG IN", currency, and language. A green navigation bar contains links for "Boston", "Hotels", "Flights", "Vacation Rentals", "Restaurants", "Things to Do", "Forum", "Best of 2016", and "More". Below this is a search bar with "Find: Hotels, Restaurants, Things to Do" and "Near: Boston, Massachusetts", with a "Search" button. The breadcrumb trail reads "United States > Massachusetts (MA) > Boston Tourism". The main heading is "Boston, Massachusetts" with "468,512 reviews and opinions" and four profile icons. A large image of the Boston skyline at night is on the left, with a "TRAVELER'S CHOICE" award badge in the top left corner. On the right, a list of categories is shown with icons, counts, review counts, and right-pointing arrows:

Category	Count	Reviews
Hotels	82	127,933 Reviews
Vacation Rentals	671	1,183 Reviews
Flights	-	-
Things to Do	761	130,851 Reviews
Restaurants	3,129	182,258 Reviews
Forum	-	25,992 Posts

Step 1

Obtain the TripAdvisor hotels page for Boston

The screenshot shows the TripAdvisor website for Boston, Massachusetts. The header includes the TripAdvisor logo, the text "Boston Tourism: Best of Boston", and links for "Review", "JOIN", "LOG IN", currency, and language. A green navigation bar contains links for "Boston", "Hotels", "Flights", "Vacation Rentals", "Restaurants", "Things to Do", "Forum", "Best of 2016", and "More". Below this is a search bar with "Find: Hotels, Restaurants, Things to Do" and "Near: Boston, Massachusetts", with a "Search" button. The breadcrumb trail reads "United States > Massachusetts (MA) > Boston Tourism". The main heading is "Boston, Massachusetts". On the left is a large image of the Boston skyline at night with a "TRAVELER'S CHOICE" award badge. On the right is a list of travel categories, with the "Hotels" category highlighted by a red box. The "Hotels" category shows "82" items and "127,933 Reviews". Other categories include "Vacation Rentals (671)", "Flights", "Things to Do (761)", "Restaurants (3,129)", and "Forum".

tripadvisor® Boston Tourism: Best of Boston Review JOIN LOG IN \$ ▼ ▼







Boston Hotels Flights Vacation Rentals Restaurants Things to Do Forum Best of 2016 More

Find: Hotels, Restaurants, Things to Do Near: Boston, Massachusetts Search

United States > Massachusetts (MA) > Boston Tourism

Boston, Massachusetts

458,543 reviews and opinions

-  **Hotels (82)**
127,933 Reviews
-  **Vacation Rentals (671)**
1,183 Reviews
-  **Flights**
-  **Things to Do (761)**
130,851 Reviews
-  **Restaurants (3,129)**
182,258 Reviews
-  **Forum**
25,992 Posts

Step 2

Extract URL of the hotels list page using BeautifulSoup

The screenshot displays the TripAdvisor website for Boston, Massachusetts. The header includes the TripAdvisor logo, navigation links (Review, JOIN, LOGIN, currency, language), and a search bar. The main content area features a large image of the Boston skyline at night and a 'TRAVELER'S CHOICE' award badge. On the right side, there is a list of travel categories, with 'Hotels (82)' highlighted by a red box. The categories listed are:

- Hotels (82) 127,933 Reviews
- Vacation Rentals (671) 1,183 Reviews
- Flights
- Things to Do (761) 130,851 Reviews
- Restaurants (3,129) 182,258 Reviews
- Forum 25,992 Posts

Step 3

Retrieve the first page of the list of hotels in Boston

The Verb Hotel

Special Offer Parking Package



[Traveler photos](#) | [Professional photos](#) | [Map](#)

03/20/2015



03/30/2015



No availability for your dates from these sites

[Expedia.com](#)

[Orbitz.com](#)

[Hotels.com](#)

[See all 9](#)

#76 Just for You

Close to Fenway Park. Offers free wifi. Other travelers love this hotel

#9 of 77 hotels in Boston

217 reviews

"What a cool hotel. Loved it!" 03/09/2015

"Great value, helpful staff, fun sta..." 03/06/2015

No Match

Days Hotel Boston ★★★★★

Special Offer Plan Ahead and Save 15%



[Traveler photos](#) | [Professional photos](#) | [Map](#)

03/20/2015



03/30/2015



No availability for your dates from these sites

[Orbitz.com](#)

[Expedia.com](#)

[Booking.com](#)

[See all 14](#)

#77 Just for You

Has a fitness center. Other travelers love this hotel. Budget hotel

#75 of 77 hotels in Boston

158 reviews

"Room was adequate" 02/13/2015

"Cheap, clean and a little walk awa..." 01/23/2015

No Match

[Previous](#)

1

2


3

[Next](#)

Step 4

Parse and extract information with BeautifulSoup

The Verb Hotel
Special Offer Parking Package


Traveler photos | Professional photos | Map

03/20/201503/30/2015

No availability for your dates from these sites

Expedia.com	❌
Orbitz.com	❌
Hotels.com	❌

See all 9

#76 Just for You

Close to Fenway Park. Offers free wifi. Other travelers love this hotel

#9 of 77 hotels in Boston


217 reviews

"What a cool hotel. Loved it!" 03/09/2015

"Great value, helpful staff, fun sta..." 03/06/2015

No Match

Days Hotel Boston ★★★★★
Special Offer Plan Ahead and Save 15%


Traveler photos | Professional photos | Map

03/20/201503/30/2015

No availability for your dates from these sites

Orbitz.com	❌
Expedia.com	❌
Booking.com	❌

See all 14

#77 Just for You

Has a fitness center. Other travelers love this hotel. Budget hotel

#75 of 77 hotels in Boston

158 reviews

"Room was adequate" 02/13/2015

"Cheap, clean and a little walk awa..." 01/23/2015

No Match

Previous

123

Next

Repeat STEPS 3 and 4 for all the other pages until the last one is reached.