

Web Scraping

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](#) [2 guests](#)

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*
\$383*

amextravel.com

\$429*

10 more sites

*Disclaimer

Luxury

Pets Allowed

Fenway / Kenmore

Traveler
photos

Professional
photos

Browse
nearby

2,916 reviews from our community

Write a Review

Traveler rating

Excellent	<div><div></div></div>	2,350
Very good	<div><div></div></div>	434
Average	<div><div></div></div>	81
Poor	<div><div></div></div>	39
Terrible	<div><div></div></div>	12

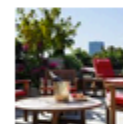
See reviews for

	Families	902
	Couples	847
	Solo	146
	Business	582

Rating summary

Sleep Quality	<div><div></div></div>
Location	<div><div></div></div>
Rooms	<div><div></div></div>
Service	<div><div></div></div>
Value	<div><div></div></div>
Cleanliness	<div><div></div></div>

Related hotels...



XV Beacon

793 Reviews

\$480 and up*



Boston Harbor Hotel

1,233 Reviews

Lowest price at Expedia, 9 sites checked

\$358 and up*



Nine Zero Hotel - a Kimpton Hotel

1,805 Reviews

Lowest price at Expedia, 8 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

\$429*

amextravel.com

\$429*

Orbitz.com

\$383*

10 more sites v

*Disclaimer

Luxury

Kenmore

Traveler
photos

Professional
photos

Browse
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



See reviews for

	Families	902
	Couples	847
	Solo	146
	Business	582

Rating summary



Related hotels...



XV Beacon

793 Reviews

\$480 and up*

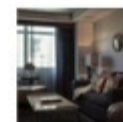


Boston Harbor Hotel

1,233 Reviews

Lowest price at Expedia, 9 sites checked

\$358 and up*



Nine Zero Hotel - a Kimpton Hotel

1,805 Reviews

Lowest price at Expedia, 8 sites checked


```

1097 </li>
1098 <li class="tabItem hvrIE6"><a onclick="ta.setEvtCookie('TopNav', 'click', 'Flights', 0, this.href);setPID(3820)" href="/Flights-g60745-Boston_Massachusetts-
Cheap_Discount_Airfares.html" class="tabLink pid4966">
1099 Flights
1100 </a>
1101 </li>
1102 <li class="tabItem hvrIE6"><a onclick="ta.setEvtCookie('TopNav', 'click', 'VacationRentals', 0, this.href)" href="/VacationRentals-g60745-Reviews-
Boston_Massachusetts-Vacation_Rentals.html" class="tabLink pid2795">
1103 Vacation Rentals
1104 </a>
1105 </li>
1106 <li class="tabItem dropDown jsNavMenu hvrIE6">
1107 <a onclick="ta.util.cookie.setPIDCookie(4967); ta.setEvtCookie('TopNav', 'click', 'Restaurants', 0, this.href)"
onmousedown="ta.common.header.addClearParam(this);" href="/Restaurants-g60745-Boston_Massachusetts.html" class="tabLink arwLink "><span
class="arrow text">Restaurants</span></a>
1108 <ul class="subNav">
1109 <li class="subItem">
1110 <a class="subLink" href="/Restaurants-g60745-Boston_Massachusetts.html" onmousedown="ta.common.header.addClearParam(this);">All Boston Restaurants</a> </li>
1111 <li class="subItem">
1112 <a class="subLink" href="/RestaurantsNear-g60745-d258705-Hotel_Commonwealth-Boston_Massachusetts.html">Restaurants near Hotel Commonwealth</a>
1113 </li>
1114 </ul>
1115 </li>
1116 <li class="tabItem dropDown jsNavMenu hvrIE6">
1117 <a onclick="ta.util.cookie.setPIDCookie(4967); ta.setEvtCookie('TopNav', 'click', 'ThingsToDo', 0, this.href)" href="/Attractions-g60745-Activities-
Boston_Massachusetts.html" class="tabLink arwLink "><span class="arrow text">Things to Do</span></a>
1118 <ul class="subNav">
1119 <li class="subItem">
1120 <a class="subLink" href="/Attractions-g60745-Activities-Boston_Massachusetts.html" onmousedown="ta.common.header.addClearParam(this);">All things to do in
Boston</a> </li>
1121 <li class="subItem">
1122 <a class="subLink" href="/AttractionsNear-g60745-d258705-Hotel_Commonwealth-Boston_Massachusetts.html">Things to do near Hotel Commonwealth</a>
1123 </li>
1124 </ul>
1125 </li>
1126 <li class="tabItem hvrIE6"><a onclick="ta.setEvtCookie('TopNav', 'click', 'TravelersChoice', 0, this.href)" href="/TravelersChoice" class="tabLink pid5087">
1127 Best of 2015
1128 </a>
1129 </li>
1130 <li class="tabItem dropDown jsNavMenu hvrIE6">
1131 <span class="tabLink arwLink"><span class="arrow text">More</span></span>
1132 <ul class="subNav">
1133 <li class="subItem">
1134 <a href="/Travel_Guide-g60745-Boston_Massachusetts.html" onclick="ta.setEvtCookie('TopNav', 'click', 'TravelGuides', 0, this.href)" class="subLink
pid16158">Travel Guides
1135 </a>
1136 </li>
1137 <li class="subItem">
1138 <a href="/ShowForum-g60745-i48-Boston_Massachusetts.html" onclick="ta.setEvtCookie('TopNav', 'click', 'TravelForum', 0, this.href)" class="subLink
pid34623">Travel Forum
1139 </a>
1140 </li>
1141 <li class="subItem">
1142 <a href="/apps" onclick="ta.setEvtCookie('TopNav', 'click', 'Apps', 0, this.href)" class="subLink pid18876">Apps
1143 </a>
1144 </li>
1145 <li class="subItem">
1146 <a href="/ShowUrl-a_partnerKey.1-a_url.http%3A_2F_2F_www_2E_cruisecritic_2E_com_2F_-a_urlKey.bb8a904288ee6bd29.html"
onclick="ta.setEvtCookie('TopNav', 'click', 'Cruises', 0, this.href)" class="subLink " target="_blank">Cruises
1147 </a>
1148 </li>
1149 <li class="subItem">
1150 <a href="/GreenLeaders" onclick="ta.setEvtCookie('TopNav', 'click', 'GreenLeaders', 0, this.href)" class="subLink pid34563">GreenLeaders
1151 </a>
1152 </li>

```

Web scraping

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)

2. Parse and extract information from the html

- html/json parsers

Retrieve Webpages

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

```
out_file = "test.html"
url = "www.tripadvisor.com"
cmd = 'curl -L -m 20 "' + url + '" > ' + ofile
os.system(cmd)
```

2. Use a python network-access libraries
 - Urllib, Request, pycurl, etc.

Parse and extract data: 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
```

Documentation available at:

<http://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Common practice

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
- **Mimic** the “*real*” http request as much as possible
 - Set HTTP headers line User-agent, Host, Referer, etc.
- Use Website **API** (if exists)

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 file 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 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:**

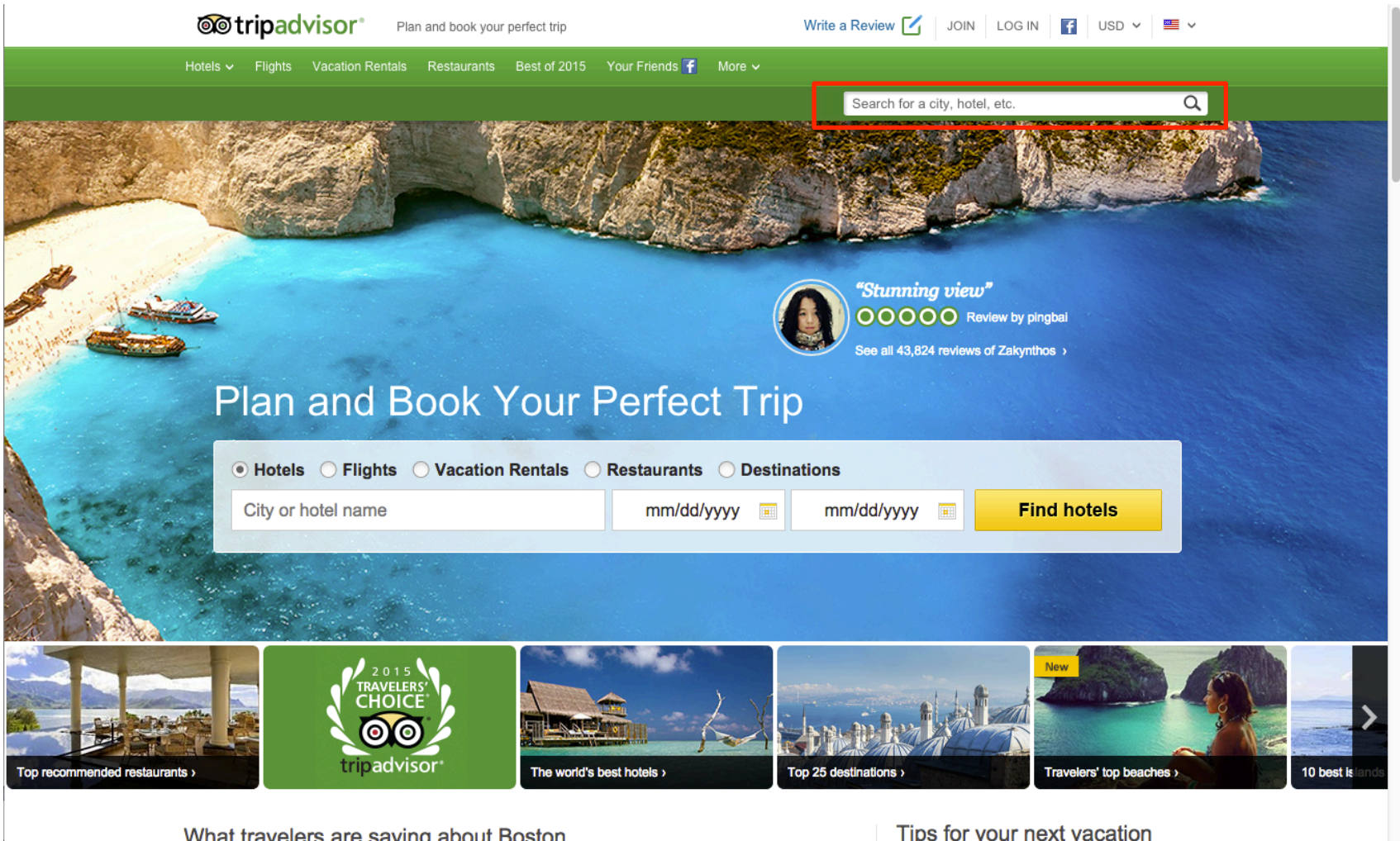
1. Use this Website or its contents for any commercial purpose
2. **Access, monitor or copy any content or information of this Website using any robot, spider, scraper or other automated means or any manual process for any purpose without our express written permission;**

Today's task

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



The screenshot shows the TripAdvisor homepage. At the top, there is a navigation bar with the TripAdvisor logo, the tagline "Plan and book your perfect trip", and links for "Write a Review", "JOIN", "LOG IN", currency selection (USD), and language selection (US flag). Below this is a green navigation menu with links for "Hotels", "Flights", "Vacation Rentals", "Restaurants", "Best of 2015", "Your Friends", and "More". A red rectangular box highlights the search bar in the top right corner of the green menu, which contains the placeholder text "Search for a city, hotel, etc." and a magnifying glass icon. The main background image is a scenic view of a beach with turquoise water and white cliffs. Overlaid on this image is a review snippet for Zakynthos, featuring a user profile picture, the text "Stunning view", five green stars, the text "Review by pingbai", and a link "See all 43,824 reviews of Zakynthos". Below the background image is a large white text overlay that reads "Plan and Book Your Perfect Trip". Underneath this is a search form with radio buttons for "Hotels", "Flights", "Vacation Rentals", "Restaurants", and "Destinations". The "Hotels" option is selected. The form includes a text input field for "City or hotel name", two date input fields labeled "mm/dd/yyyy" with calendar icons, and a yellow "Find hotels" button. At the bottom of the page, there is a horizontal row of six promotional tiles: "Top recommended restaurants", "2015 TRAVELERS' CHOICE award" (with the TripAdvisor owl logo), "The world's best hotels", "Top 25 destinations", "Travelers' top beaches" (marked with a "New" tag), and "10 best islands". Each tile has a right-pointing arrow. Below the tiles, there are two links: "What travelers are saving about Boston" and "Tips for your next vacation".

tripadvisor® Plan and book your perfect trip

Write a Review JOIN LOG IN USD US

Hotels Flights Vacation Rentals Restaurants Best of 2015 Your Friends More

Search for a city, hotel, etc.

"Stunning view"
★★★★★ Review by pingbai
See all 43,824 reviews of Zakynthos

Plan and Book Your Perfect Trip

☒ Hotels ☐ Flights ☐ Vacation Rentals ☐ Restaurants ☐ Destinations

City or hotel name mm/dd/yyyy mm/dd/yyyy **Find hotels**

Top recommended restaurants

2015 TRAVELERS' CHOICE
tripadvisor®

The world's best hotels

Top 25 destinations

New
Travelers' top beaches

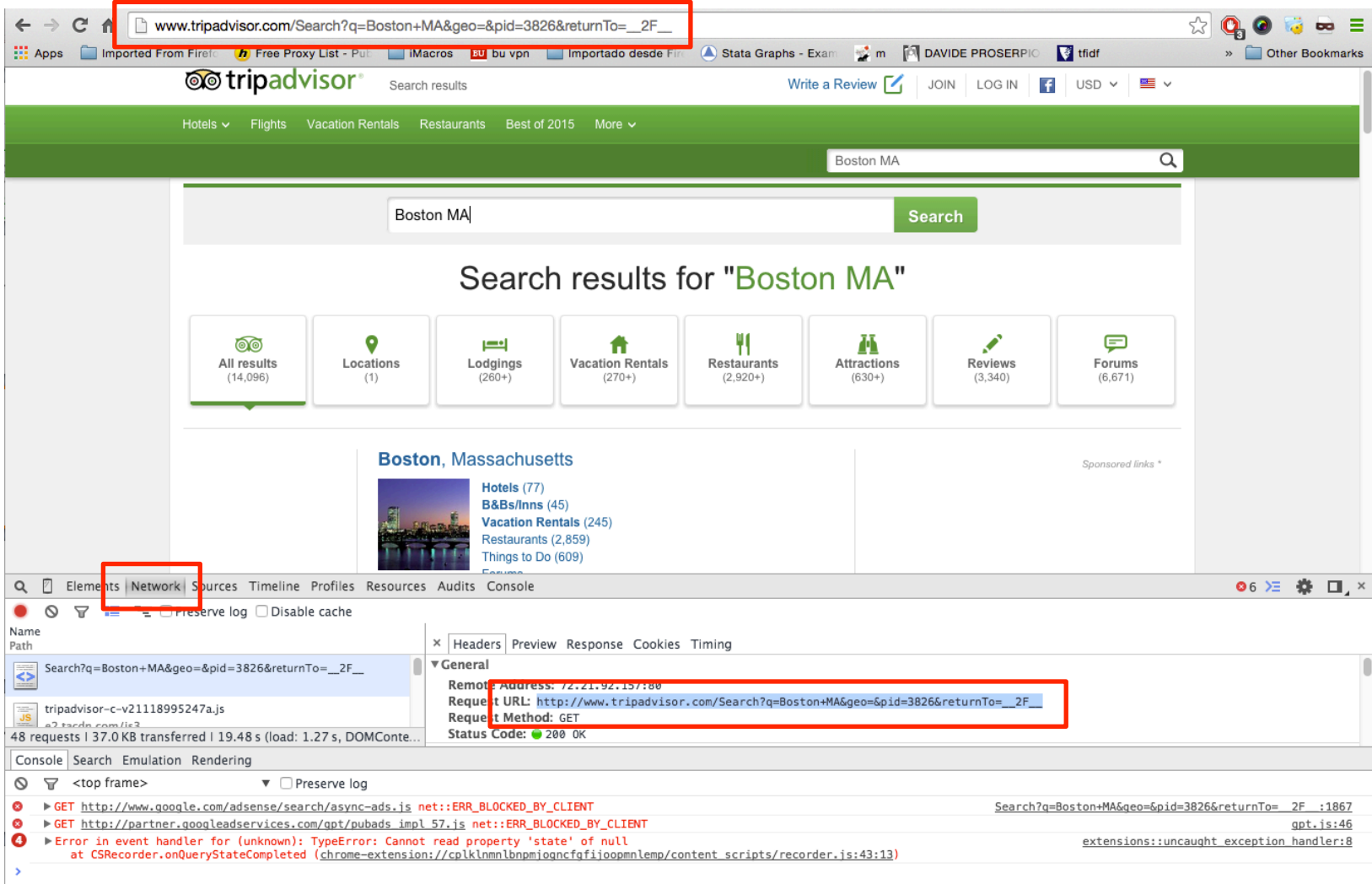
10 best islands

What travelers are saving about Boston

Tips for your next vacation

Step 1

Obtain the TripAdvisor Boston page



Step 1

Obtain the TripAdvisor Boston page



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 "Write a Review", "JOIN", "LOG IN", "USD", and a flag icon. A green navigation bar contains links for "Boston", "Hotels", "Flights", "Vacation Rentals", "Restaurants", "Things to Do", "Best of 2015", and "More". Below this is a search bar with the text "Search for a city, hotel, etc." and a magnifying glass icon.

The main content area features a large banner image of the Boston skyline at night with the text "TripAdvisor Boston Massachusetts" and "279,763 reviews and opinions". Below the banner are four smaller images: "19,699 candid traveler photos", a video player with "BOSTON" text, a city skyline, and a "TRAVELLER'S CHOICE" award logo.

On the right side, there is a vertical list of categories with icons and counts:

- Hotels (77) 86,097 Reviews
- Vacation Rentals (251) 1,229 Reviews
- Flights
- Attractions (609) 65,581 Reviews
- Restaurants (2,858) 103,818 Reviews
- Forum 22,744 Posts
- Travel Guides 6 Guides

Below the banner, there is a section titled "Walk the Freedom Trail the first time you visit Boston, and you'll quickly get a sense of this coastal city's revolutionary spirit and history. But make sure you also explore some of Boston's fine museums (try the Gardner, featuring art masterpieces displayed in their collector's mansion) and old neighborhoods (like the North End, Boston's Little Italy). You can't..." with a "Read more" link.

Next to this text is a circular collage of images with the text "See what your friends say about Boston" and a "Log in with Facebook" button.

At the bottom right, there is a section titled "Don't miss the best of Boston" with two images: "3 Days in Boston" and "Guide to Boston Outdoors". Below these images is a link "See all travel guides".

Step 2

Extract URL of the hotels list page using BeautifulSoup

The screenshot displays the TripAdvisor website for Boston, Massachusetts. The top navigation bar includes the TripAdvisor logo, the location "Boston Tourism: Best of Boston", and links for "Write a Review", "JOIN", "LOG IN", and currency/language settings. Below this is a green navigation bar with tabs for "Boston", "Hotels", "Flights", "Vacation Rentals", "Restaurants", "Things to Do", "Best of 2015", and "More". A search bar is located on the right side of this bar.

The main content area features a large banner image of the Boston skyline at night. Below the banner, there are several sections: "279,763 reviews and opinions", "19,699 candid traveler photos", a "BOSTON" video player, and a "TRAVELLER'S CHOICE" award badge. To the right of the banner is a sidebar with a list of travel categories, each with an icon, a count, and a link to reviews. The "Hotels (77)" category is highlighted with a red box.

Below the main content area, there is a section titled "Walk the Freedom Trail the first time you visit Boston, and you'll quickly get a sense of this coastal city's revolutionary spirit and history. But make sure you also explore some of Boston's fine museums (try the Gardner, featuring art masterpieces displayed in their collector's mansion) and old neighborhoods (like the North End, Boston's Little Italy). You can't..." with a "Read more" link. To the right of this text is a "See what your friends say about Boston" section with a "Log in with Facebook" button. Further right is a "Don't miss the best of Boston" section with two featured travel guides: "3 Days in Boston" and "Guide to Boston Outdoors". A "See all travel guides" link is located below these guides.

Category	Count	Reviews
Hotels	77	86,097 Reviews
Vacation Rentals	251	1,229 Reviews
Flights		
Attractions	609	65,581 Reviews
Restaurants	2,858	103,818 Reviews
Forum		22,744 Posts
Travel Guides	6	6 Guides

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

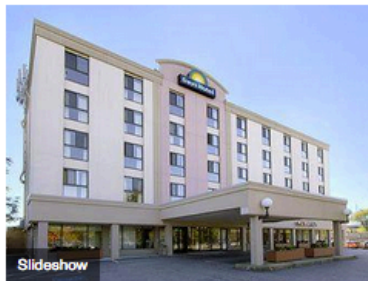
"What a cool hotel. Loved it!" 03/09/2015

"Great vibe, helpful staff, fun sta..." 03/09/2015

 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

#76 of 77 hotels in Boston

 188 reviews

"Better than 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


Slideshow
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 vibe, helpful staff, fun sta..." 03/09/2015

No Match

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


Slideshow
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

#76 of 77 hotels in Boston

188 reviews

"Better than 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.