



CS 329P : Practical Machine Learning (2021 Fall)

## 1.3 Web Scraping



Qingqing Huang, Mu Li, Alex Smola

<https://c.d2l.ai/stanford-cs329p>

# Web Scrapping




- The goal is to extract data from website
  - Noisy, weak labels, can be spammy
  - Available at scale
  - E.g. price comparison/tracking website
- Many ML datasets are obtained by web scraping
  - E.g. ImageNet, Kinetics
- Web crawling VS scrapping
  - Crawling: indexing whole pages on Internet
  - Scrapping: scraping particular data from web pages of a website



Image credit: Aaron Zappia

# Tools



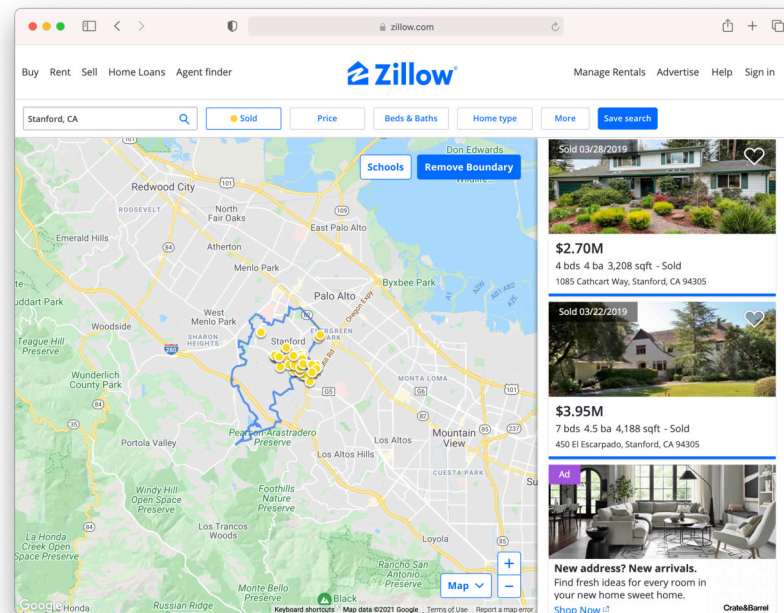
- “curl” often doesn’t work 
  - Website owners use various ways to stop bots
- Use headless browser: a web browser without a GUI
- You need a lot of new IPs, easy to get through public clouds
  - In all IPv4 IPs, AWS owns 1.75%, Azure 0.55%, GCP 0.25%

```
1 from selenium import webdriver
2
3 chrome_options = webdriver.ChromeOptions()
4 chrome_options.headless = True
5 chrome = webdriver.Chrome(
6     chrome_options=chrome_options)
7
8 page = chrome.get(url)
```

# Case Study



- Query houses sold in near Stanford
- <https://www.zillow.com/stanford-ca/sold/>
- <https://www.zillow.com/stanford-ca/sold/2-p/>
- ...
- You can replace the city and state in the URL for other places



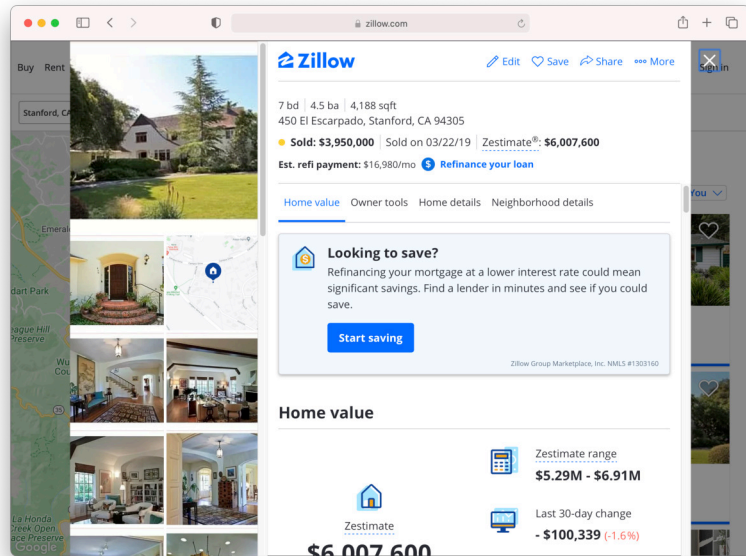
# Crawl individual pages



- Get the house IDs from the index pages

```
1 page = BeautifulSoup(open(html_path, 'r'))
2 links = [a['href'] for a in page.find_all(
3     'a', 'list-card-link')]
4 ids = [l.split('/')[2].split('_')[0]
5     for l in links]
```

- The house detail page by ID
  - [https://www.zillow.com/homedetails/19506780\\_zpid/](https://www.zillow.com/homedetails/19506780_zpid/)



# Extract data



- Identify the HTML elements through  
**Inspect**

```
1 sold_items = [a.text for a in page.find(
2     'div', 'ds-home-details-chip').find('p').find_all('span')]
3 for item in sold_items:
4     if 'Sold:' in item:
5         result['Sold Price'] = item.split(' ')[1]
6     if 'Sold on' in item:
7         result['Sold On'] = item.split(' ')[-1]
```



450 El Escarpado, Stanford, CA

zillow.com/homedetails/450-El-Escarpado-Stanford-CA-94305/19506780\_zpid/

Incognito

Update

1 of 24

span.sc-pZcYF.kUpbpZ.ds-status-details 142.8 x 22

Sold: \$3,950,000 Sold on 03/22/19

Zestimate®: \$6,007,600

Est. refi payment: \$16,969/mo Refinance your loan

Home value Owner tools Home details Neighborhood de

Elements Sources Console Network

8-48-0\_\_sc-ai24-0 StyledHeading-c11n-8-48-0\_\_sc-ktujwe-0 YmdCA>...</h1>

</div>

</div>

<p class="Text-c11n-8-48-0\_\_sc-ai24-0 StyledParagraph-c11n-8-48-0\_\_sc-18ze78a-0 ghsFfg">

><span class="sc-pZcYF kUpbpZ ds-status-details">...

</span> == \$0

><span class="sc-oTzDS fotNM">...</span>

><span class="sc-oTzDS fotNM">...</span>

</p>

<div class="sc-prqHV gZvZRY ds-chip-removable-content">...</div>

><div class="ds-mortgage-row">...</div>

<div class="sc-pBAKv hDPbjp">...</div>

</div>

</div>

...-0.StyledParagraph-c11n-8-48-0\_\_sc-18ze78a-0.ghsFfg span.sc-pZcYF.kUpbpZ.ds-status-details

# Extract data

- Repeat the previous process to extract other field data



**Zillow** [Save](#) [Share](#) [More](#)

4 bd | 4 ba | 2,939 sqft  
44626 Sandia Creek Dr, Temecula, CA 92590  
**Sold: \$960,000** | Sold on 02/11/21 | Zestimate®: **\$1,119,900**  
Est. refi payment: \$4,308/mo [Refinance your loan](#)

Home value Owner tools **Home details** Neighborhood details

### Overview

This single level home on 9.66 acres land is beautifully maintained and has a lot of features you don't find in a typical tract home! there is a guest home of aprox. 566 s.f. (included in the overall square footage) attached to a 4-car garage! great picturesque views of the hills and valleys of the De Luz Custom Home community in the hills to the West of the City of Temecula! Enjoy breathtaking sunsets off a big balcony that features trex decking. main house features 3 bed and 2.5 baths with wood flooring

Listed by:  
George Tektonopoulos DRE# 00999662  
Tekton Real Estate

Source: SDMLS, MLS#: SW21002332 **SAN DIEGO** | MLS

Zillow checked: September 09, 2021 at 10:32pm  
Data updated: February 12, 2021 at 12:45pm

Bought with: Marcie George  
First Team Real Estate

### Facts and features [Edit](#)

<b>Type:</b> Detached	<b>Parking:</b> 4 Garage spaces
<b>Year built:</b> 1986	<b>Lot:</b> 9.66 Acres
<b>Heating:</b> Forced Air	<b>Buyer's Agent Fee:</b> 2.5%
<b>Cooling:</b> Central Forced Air	

### Interior details

Bedrooms and bathrooms <ul style="list-style-type: none"><li>Bedrooms: 4</li></ul>	Interior Features <ul style="list-style-type: none"><li>Interior features: Bedroom Entry</li></ul>
--	--

[See more facts and features](#)

### Services availability

Zillow Internet Resource Center  
[See Providers](#) | [Compare Speeds](#) | [Get Deals](#)

### Price and tax history

#### Price history



```
1 { 'Id': '18173100',
2   'Address': '44626 Sandia Creek Dr,',
3   'Sold Price': '$960,000',
4   'Sold On': '02/11/21',
5   'Summary': "This single level home on 9.66 acres la
6   'Type': 'SingleFamily',
7   'Year built': '1986',
8   'HOA': '$70 annually',
9   'Lot': '420,790 sqft',
10  'Bedrooms': '4',
11  'Bathrooms': '4',
12  'Full bathrooms': '3',
13  '1/2 bathrooms': '1',
14  'Main level bathrooms': '3',
15  'Association name': 'De Luz Ranchose',
16  'Tax assessed value': '$784,118',
17  'Annual tax amount': '$12,036',
18  'Listed On': '1/5/2021',
19  'Listed Price': '$975,000',
20  'Last Sold On': '3/5/2008',
21  'Last Sold Price': '$840,000',
22  'Elementary School': 'Murrieta Elementary School',
23  'Elementary School Score': '7',
24  'Elementary School Distance': '4.9',
25  'Middle School': 'Thompson Middle School',
26  'Middle School Score': '6',
27  'Middle School Distance': '5.5',
28  'High School': 'Murrieta Valley High School',
29  'High School Score': '8',
```

# Cost



- Use AWS EC2 t3.small (2GB memory, 2 vCPUs, \$0.02 per hour)
  - 2GB is necessary as the browser needs a lot memory, CPU and bandwidth are usually not an issue
  - Can use spot instance to reduce the price
- The cost to crawl 1M houses is \$16.6
  - The speed is about 3s per page,
  - 8.3 hours if using 100 instances
  - The extra cost includes storage, restart instances when IP is banned





# Crawl Images



- Get all image URLs




```
1 p = r'https:\\\\photos.zillowstatic.com\\fp\\([\\d\\w\\-\\_]+).jpg'  
2 ids = [a.split('-')[0] for a in re.findall(p, html)]  
3 urls = [f'https://photos.zillowstatic.com/fp/{id}-  
uncropped_scaled_within_1536_1152.jpg' for id in ids]
```

- A house listing has ~20 images
  - The crawling cost is still reasonable: ~\$300
  - Storing these images is expensive: ~\$300 per month
    - You can reduce the image resolutions, or send data back

# Legal Considerations



- Web scraping isn't illegal by itself
- But you should 
  - NOT scrape data have sensitive information (E.g. private data involving username/password, personal health/medical information)
  - NOT scrape copyrighted data (E.g. YouTube videos, Flickr photos)
  - Follow the Terms of Service that explicitly prohibits web scraping
- Consult a lawyer if you are doing it for profit

# Summary



- Web scraping is a powerful way to collect data at scale when the website doesn't offer a data API
- Low cost if using public clouds
- Use browser's inspection tool to locate the information in HTML
- Be cautious to use it properly