



Web Scraping Project



Scraping Hotels list from Booking.Com



Overview

- Objective & Tools
- Preliminary Details
- Experimentation Phase
- Scraping & Saving
- Summary and Limitations

Objectives & Tools

Objectives

- Pick out a fixed location
- Make a list of all available hotels.
- Information: Name, Location, Reviews etc.
- Make a csv file for further analysis

Tools

- Website: Booking.Com
- Jupyter Notebook
- Python 3
- Python Libraries: bs4, pandas, lxml, requests & jovian

Special Note:

- *Web Scraping Projects are highly prone to error and malfunction.*
- *Check beforehand for the consent, permissions and privacy policy.*

Preliminary Details

Search

Destination/property name:

 New Delhi

Check-in date



Check-in Date



Check-out date



Check-out Date



2 adults



No children



1 room

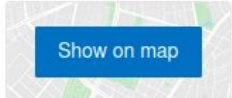


I'm traveling for work



Search

New Delhi: 1,742 properties found

 Show on map

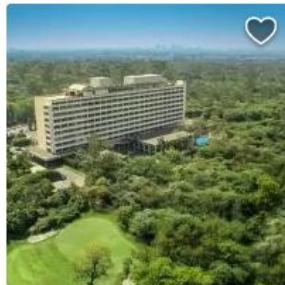
Our Top Picks

Entire homes & apartments

Stars (highest first)

Stars (lowest first)

Top Reviewed



The Oberoi New Delhi ★★★★★

[New Delhi](#) · [Show on map](#) · 4.6 km from center

Overlooking Delhi's Golf Club and located in Central Delhi, newly renovated The Oberoi, New Delhi is 9.9 mi from the Indra Gandhi International Airport.

Exceptional

460 reviews

9.6

Location 9.5

Show prices



Bungalow 99

[South Delhi, New Delhi](#) · [Show on map](#) · 6.8 km from center

Exceptional

216 reviews

9.7

Preliminary Information:

- Location : New Delhi
- Category : Top Reviewed
- Page Results : 25 Posts Per Page ($25 \times 40 = 1000$)
- Links : First Page & Last Stage

Scraping Content

New Delhi: 1,742 properties found

Show on map

Our Top Picks

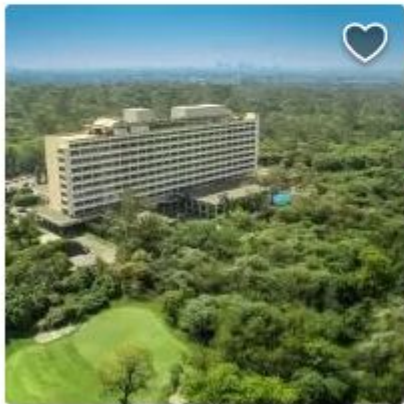
Entire homes & apartments

Stars (highest first)

Stars (lowest first)

Top Reviewed

...



The Oberoi New Delhi ★★★★★

[New Delhi](#) • [Show on map](#) • 4.6 km from center

Overlooking Delhi's Golf Club and located in Central Delhi, newly renovated The Oberoi, New Delhi is 9.9 mi from the Indira Gandhi International Airport.

Exceptional **9.6**
460 reviews

Location 9.5

Show prices

- Hotel Name
- Location
- Reviews: Rating, Category, Counts
- Distance (From City Center)
- Metro Facilities
- Description

Sample Format

```
# To see the csv file  
sample_df
```

	Hotel Name	Place	Review (Points)	Review (Cat.)	Review (Count)	Distance (City Centre)	Metro Facilities	Description
0	The Oberoi New Delhi	New Delhi	9.6	Exceptional	460 reviews	4.6 km from centre	No Metro Access	Overlooking Delhi's Golf Club and located in C...
1	Bungalow 99	South Delhi, New Delhi	9.7	Exceptional	216 reviews	6.8 km from centre	No Metro Access	Located in New Delhi, 2.6 mi from Humayun's To...

Experimentation Phase

Fix Browsing Details:

- Main Link
- Headers : As per OS and System

```
: main_link = "https://www.booking.com/searchresults.html?aid=304142&label=gen173nr-1FCAEoggI46AdIM1gEaGyIAQ6QH4AQKIAgGoAg04AsrBzIcGwAIB0gIkNDFiZWU4NjMtMDU1My00NjA4LTkxMmYtMGE2NzliNTRiNTNl2AIF4AIB&sid=55fa6423aa50269tpl=searchresults&city=-2106102&class_interval=1&dest_id=-2106102&dest_type=city&dr_ps=IDR&from_idr=1&groupchildren=0&ilp=1&label_click=undef&no_rooms=1&order=bayesian_review_score&raw_dest_type=city&room1=A%2CA&sb_hw_aparth=1&slp_r_match=0&srpvid=518283f1feb10049&ssb=empty&top_ufig=1&sig=v1Q1dbuDSb&rows=25&offset=0"

headers = {'User-Agent': 'Mozilla/5.0 (platform; rv:geckoversion) Gecko/geckotrail Firefox/firefoxversion'}
```

Testing Phase

```
: # Request for the URL  
page = requests.get(main_link, headers=headers)  
  
# We can also check the response code  
print(page.status_code)  
  
# Make it a soup  
soup = BeautifulSoup(page.text, "lxml")  
  
# Display Soup ( Main Page)  
soup  
  
200  
: <!DOCTYPE html>
```

Section Search

- Inspect Elements
- Look for distinguishable sections
- Search for Class, Anchor, Div, Paragraph Tag & more

Search

Destination/property name:

🔍 New Delhi

Check-in date

📅 Check-in Date

Check-out date

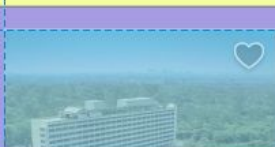
📅 Check-out Date

New Delhi: 1,742 properties found

Show on map

Our Top Pick

div.sr_item.sr_item_new.sr_item_default.sr_property_block.sr_flex_layout.sr_item_no_dates 814 × 234 Flex Container



The Oberoi New Delhi ★★★★★

New Delhi · Show on map · 4.6 km from center

Overlooking Delhi's Golf Club and located in Central Delhi, newly renovated
The Oberoi, New Delhi is 9.9 mi from the Indira Gandhi International Airport.

Exceptional
460 reviews
Location 9.5

Search HTML

```
<div id="hotellist_inner" class="nodates_hotels wider_image" role="region"
aria-label="Search results" data-et-view=" OQLOLOMSVSRqbVKKMadMUWe:1
OQLOLOLOVCJVIQNCgzdMbHGQWKHC:2 ... OLGPNCeIWDAdbIVXVEHINET:6
OLGPNCeIWDAdbIVXVEHINET:8 ">
  <div class="sr_item sr_item_new sr_item_default sr_property_block
sr_flex_layout sr_item_no_dates " data-et-
click="customGoal:NAREFBINEIfBccOHT:2" data-hotelid="77814" data-class="5"
data-score="9.6"> flex
    <div id="hotel_77814" class="sr_item_photo sr_card_photo_wrapper"> </div>
    <div class="sr_item_content sr_item_content_slider_wrapper ">
      <div class="sr_property_block_main_row"> flex
        <div class="sr_item_main_block">
          <div class="sr_hotel_title_wrap">
            <h3 class="sr_hotel_title ">
              <a class="is-sr_hotel.link_hotel_name link_url" href="/hotel
div.sr_item.sr_item_new.sr_item_default... div.sr_item_content.sr_item_content_slid... div.sr_property_block_main
```

Filter Styles

:hov .cls + 🌞 🌙 📄

```
element {
}
9dd837928987b5cfa928a864f74604fe7b2e5716.css:1
.sr_flex_layout.sr_item {
  -webkit-box-align: top;
  -webkit-align-items: top;
  -ms-flex-align: top;
  align-items: top;
  display: -webkit-box;
  display: -ms-flexbox;
  display: flex;
  -webkit-box-orient: horizontal;
  -webkit-box-direction: normal;
  -webkit-flex-direction: row;
  -ms-flex-direction: row;
```

Layout Computed Changes Fonts Anim

Flex Container

div.sr_item.sr_item_new.sr_item_def... row wrap

Flex Items

- 1 div.clearfix
- 2 div#hotel_77814.sr_item_photo.sr_card_ph...
- 3 div.sr_item_content.sr_item_content_slide...

Grid

CSS Grid is not in use on this page

```
<span class="invisible_spoken">Opens in new window</span>
</a>
</td>
<td>
<div class="sr-hotel_title-wrap">
<h3 class="sr-hotel_title">
<a class="js-sr-hotel-link hotel_name_link url" href="
/hotel/in/the-oberoi-new-delhi.en-gb.html?aid=304142&label=gen173nr-1FCAEoggI46AdIMlgEaGyIAQGYATG4ARnIAQ_YAQH
oAQH4AQKIAGoAg04AsrBzIcGwAIB0gIkNDFiZWU4NjMtMDU1My00NjA4LTkxMmYtMGE2NzliNTRiNTNl2AIF4AIB&sid=0a70de659675e2e
2523032ae2a4118e3&dest_id=-2106102&dest_type=city&group_adults=2&group_children=0&hapos=1&hpos=1&no_rooms=1&sr_order=
bayesian_review_score&srepoch=1626597664&srpvid=7ae53d0fa8410132&ucfs=1&sig=v13_qVGitz&from=searchresults
;highlight_room=#hotelTpl" rel="noopener" target=" blank">
<span class="sr-hotel_name" data-et-click=" ">
The Oberoi New Delhi
</span>
<span class="invisible_spoken">Opens in new window</span>
</a>
</h3>
<span class="sr-hotel_title-badaes">
```

```
In [13]: # Select the First Section
section = soup.select(".sr_item.sr_item_new.sr_item_default.sr_property_block.sr_item_no_dates")[0]

# Select the second anchor tag
anchor = section.select("a")[1]

# Select the first span tag
span = anchor.select("span")[0]

# Get the name of hotel
name = span.getText().split('\n')[1]
name
```

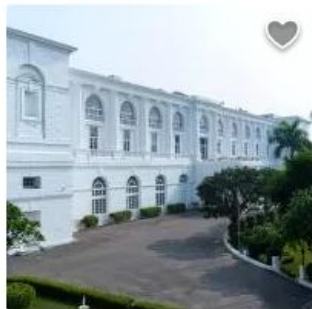
```
Out[13]: 'The Oberoi New Delhi'
```

Building Functions

- Easy to use functions
- Proper Output
- Add-On Notes

```
In [14]: def get_hotel_name(section):  
         """  
         Function gives us the hotel name  
         """  
         anchor = section.select("a")[1]  
         span = anchor.select("span")[0]  
         name = span.getText().split('\n')[1]  
         return name
```

Error Handling



Maidens Hotel New Delhi ★★★★★

[North Delhi, New Delhi](#) · [Show on map](#) · 4.4 km from center

· [Subway Access](#)

Built in 1903, Maidens Hotel showcases 19th century colonial charm and architecture. It has an outdoor pool, fitness center and features a coffee shop which extends into a charming, open courtyard.

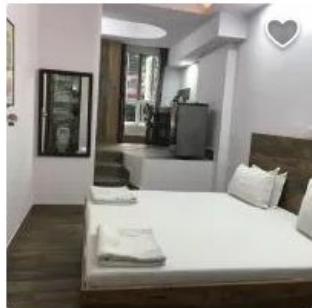
Booked 3 times in the last 24 hours

Wonderful

762 reviews

9.4

Show prices



Tatvamasi Homestay

[South Delhi, New Delhi](#) · [Show on map](#) · 11.2 km from center

Tatvamasi Homestay is located in New Delhi. It is 1 mi from the serene Lotus Temple. Free WiFi access is available.

Exceptional

118 reviews

9.8

Location 9.7

Show prices

```
In [26]: def get_distance_info(section):  
    """  
    Function returns the distance information  
    """  
    # For the centre distance  
    center_distance = section.select(".sr_card_address_line__user_destination_address")[0].getText()  
    distance = center_distance.split("\n")[1]  
  
    # For Metro Access  
    try:  
        metro_list = section.select(".sr_card_address_line__item")[1].getText()  
        metro = metro_list.split("\n")[2]  
    except:  
        metro = 'No Metro Access'  
    return distance, metro
```

Scraping & Saving

Final Stage

- Build a Dictionary for saving data
- Make a Composite Functions
- Execution and Wait

```
# Setting up the URL and headers
url = "https://www.booking.com/searchresults.html?aid=304142&label=gen173nr-1FCAEoggI46AdIM1gEaGyIAQGYATG4ARnIA
headers = {'User-Agent': 'Mozilla/5.0 (platform; rv:geckoversion) Gecko/geckotrail Firefox/firefoxversion'}

# Creating empty dictionary for storing
scrape_dict = new_data_dict()

# Looping through all the content from 0-1000
for num in range(0,976,25):

    # Request for the URL and make soup
    soup = get_page(url,headers,num)

    # Get the sections
    sections = soup.select(".sr_item.sr_item_new.sr_item_default.sr_property_block.sr_item_no_dates")

    # For all the blocks
    for section in sections:

        # Get and add the details to dictionary
        scrape_dict["Name"].append(get_hotel_name(section))
        scrape_dict["Place"].append(get_location(section))
        part1,part2,part3 = get_review_details(section)
        scrape_dict["Rating Point"].append(part1)
        scrape_dict["Rating Category"].append(part2)
        scrape_dict["Number of Reviews"].append(part3)
        info1,info2 = get_distance_info(section)
        scrape_dict["Distance from Centre"].append(info1)
        scrape_dict["Metro Access"].append(info2)
        scrape_dict["Description"].append(get_description(section))
```

View & Save

In [37]: df[:20]

Out[37]:

	Name	Place	Rating Point	Rating Category	Number of Reviews	Distance from Centre	Metro Access	Description
0	The Oberoi New Delhi	New Delhi	9.6	Exceptional	460 reviews	4.6 km from centre	No Metro Access	Overlooking Delhi's Golf Club and situated in ...
1	Bungalow 99	South Delhi, New Delhi	9.7	Exceptional	216 reviews	6.8 km from centre	No Metro Access	Situated in New Delhi, 4.2 km from Humayun's T...
2	The Leela Palace New Delhi	South Delhi, New Delhi	9.3	Superb	1,240 reviews	7 km from centre	No Metro Access	Located in New Delhi's Diplomatic Enclave, The...
3	Maidens Hotel New Delhi	North Delhi, New Delhi	9.4	Superb	762 reviews	4.4 km from centre	Metro access	Built in 1903, Maidens Hotel showcases 19th ce...
4	Tatvamasi Homestay	South Delhi, New Delhi	9.8	Exceptional	118 reviews	11.2 km from centre	No Metro Access	Tatvamasi Homestay is located in New Delhi. It...
5	The Imperial, New Delhi	Connaught Place, New Delhi	9.3	Superb	791 reviews	1.3 km from centre	Metro access	Located 1 km from New Delhi's City Centre and ...
6	Cp Villa	Central Delhi, New Delhi	9.3	Superb	329 reviews	1.5 km from centre	Metro access	Featuring free WiFi, Cp Villa is situated in N...
7	Tobo Stays	South Delhi, New Delhi	9.2	Superb	376 reviews	12.1 km from centre	No Metro Access	Situated in New Delhi and with Lotus Temple re...
8	Taj Palace, New Delhi	Chanakyapuri, New Delhi	8.9	Fabulous	3,379 reviews	6.7 km from centre	No Metro Access	Spread over six acres of lush gardens, offerin...
9	Staybook - ShivDev	Bahamani, New Delhi	8.5	Very good	848 reviews	1.3 km from	No Metro	Staybook - ShivDev

Summary, Limitations & Future Work

Summary

- We scraped the data of hotels in New Delhi city from Booking.com Website.
- We build a comprehensive list of 1000 hotels
- Scraped details like name, location, reviews, metro access, distance from city centre and description.

Limitations

- The maximum number of posts we can see is 1000, so we were only able to scrape data of 1000 out of 1700 available hotels.
- Due to constant call-return of url and results, there may be some duplicates.
- Not all hotels are available all the time, so it's highly probable that we missed out some.

Future Work

- We can try to scrape further details related to the hotel like its facilities, images and all.
- We can build a bigger project consisting of list multiple cities.
- We can make our code better by adding features like error handling, duplicates removal and more.



Thank You



Bhupesh Chandrawanshi

