



SISTEMI INFORMATIVI E DBMS

CORSO DI LAUREA MAGISTRALE IN
MANAGEMENT E MONITORAGGIO DEL TURISMO SOSTENIBILE

ESERCITAZIONE 5

Tool di Web Scraping

Ing. Gavina Baralla

Web scraping

Tecnica informatica che consente l'estrazione dati da un sito web.

Nota anche come **Web harvesting** o **Web data extraction**.

L'estrazione dei dati può essere fatta utilizzando programmi software già disponibili oppure implementando del codice ad hoc.

Esempio di web scraping: estrapolare recensioni

Cosa fare e vedere a Buggerru

Quali sono le date del viaggio?


[Le migliori attività](#) [Tour e biglietti](#)


Esplora per categoria

[Spiagge](#) [Parchi e natura](#) [Attività all'aperto](#)


Le principali attrazioni a Buggerru


Ordina per: [Preferiti dai viaggiatori](#)


- 


Cala Domestica
508 recensioni
- 


Galleria Henry Buggerru
258 recensioni


- 


Spiaggia di San Nicolo'
21 recensioni
- 


Dido Beach
43 recensioni
- 

Ondanomala Portixeddu
27 recensioni
- 

Spiaggia di BUGGERRU
0 recensioni
- 

Spiaggia di Buggerru
8 recensioni
- 

Museo del Minatore di Buggerru
12 recensioni
- 

The smile of summer
3 recensioni
- 

Oasi Beach Tours
12 recensioni

[Scopri di più](#)

Implementazione di script ad hoc

Libreria Python

Beautiful Soup

```
import bs4
from urllib.request import urlopen as uReq
from bs4 import BeautifulSoup as soup

url = 'https://www.tripadvisor.it/Attractions-g1485417-Activities-B

uClient = uReq(url)
page_html = uClient.read()
uClient.close()

filename = "attrazioni_bugerru.csv"
f = open(filename, "w", encoding = "utf-8")

page_soup = soup(page_html, "html.parser")

containers = page_soup.findAll("div", {"class": "attractions-attracti
containers2 = page_soup.findAll("div", {"class": "attractions-attract
containersT = page_soup.findAll("h1")

print(containersT[1].text.strip())
f.write('\n' + containersT[1].text.strip() + '\n')

for link2 in containers:
    link2 = link2.a

    url2 = 'https://www.tripadvisor.it'+ link2['href']

    uClient = uReq(url2)
    page_html = uClient.read()
    uClient.close()

    page_soup = soup(page_html, "html.parser")

    containers = page_soup.findAll("p", {"class": "partial_entry"})
    containersT = page_soup.find("h1", {"class": "ui_header h1"})
    containers = containers[2:]

    for container in containers:
        if containersT != None:
            #title = container.div.text.strip()
            post = container.text.strip()
            f.write('\n' + containersT.text.strip() + ',' + post + "\n")

for container in containers:
    if containersT != None:
        #title = container.div.text.strip()
        post = container.text.strip()
        f.write('\n' + containersT.text.strip() + ',' + post + "\n")

for link in containers2:
    link = link.a
    url = 'https://www.tripadvisor.it'+ link['href']

    uClient = uReq(url)
    page_html = uClient.read()
    uClient.close()

    page_soup = soup(page_html, "html.parser")

    containers = page_soup.findAll("p", {"class": "partial_entry"})
    containersT = page_soup.find("h1", {"class": "ui_header h1"})
    containers = containers[2:]

    for container in containers:
        if containersT != None:
            #title = container.div.text.strip()
            post = container.text.strip()
            f.write('\n' + containersT.text.strip() + ',' + post + "\n")
```

Tool di Web Scraping

botsol

<http://www.botsol.com/>

 | Octoparse

<https://www.octoparse.com/>

parsehub

<https://www.parsehub.com/>

Botsol: Google Maps Crawler

Botsol Crawler | Select Bot

Please select the bot you want to run.

- Gmaps Reviews Crawler V5
- Google Maps Crawler Free
- Yellowpages Crawler Free
- Amazon Product Data Crawler Free
- Yelp Crawler Free

The screenshot shows a Google Maps search for 'restaurants in Toronto, ON, Canada'. The search results list several restaurants, including Cafe Diplomatico Restaurant & Pizzeria, Canoe, and George Restaurant. A Botsol crawler window is overlaid on the map, displaying instructions and a table of extracted data.

Botsol: Google Maps Crawler Free : Botsol app v7.0

Instructions: The bot will open a chrome window, use that window to search for businesses and when search results are shown click on the "Start Bot" button to start the crawler. Make sure language for google maps is set to English (United States) Do not click on the close or minimize buttons of chrome

Progress: Record # 14 ----- Name:Marben Full Address:488 Wellington St W, Toronto, ON M5V 1E3, Canada Website:marben.ca

Name	Full Address	Website	Plus Code	Rating	Reviews	URL
Canoe	66 Wellington St W, Toronto, ON M5V 1E3, Canada	canoerestaurant.com	JJX9+2J Toronto, ON	4.4	1,101 reviews	https://www.google.com/maps/place/Canoe+Restaurant+Pizzeria/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s66+Wellington+St+W,+Toronto,+ON+M5V+1E3,+Canada
George Restaurant	111C Queen St E, Toronto, ON M5C 1S6, Canada	georgeonqueen.com	MJ3G+85 Toronto, ON	4.6	482 reviews	https://www.google.com/maps/place/George+Restaurant/@43.6530556,-79.3719444,15z/data=!3m1!1e3!3m2!1s111C+Queen+St+E,+Toronto,+ON+M5C+1S6,+Canada
Richmond Station	1 Richmond St W, Toronto, ON M5H 2L4, Canada	richmondstation.ca	MJ2C+H6 Toronto, ON	4.7	1,144 reviews	https://www.google.com/maps/place/Richmond+Station/@43.64749,-79.3719444,15z/data=!3m1!1e3!3m2!1s1+Richmond+St+W,+Toronto,+ON+M5H+2L4,+Canada
The Elm Tree	43 Elm St, Toronto, ON M5R 1A5, Canada	theelmtree.ca	MJ48+VG Toronto, ON	4.3	348 reviews	https://www.google.com/maps/place/The+Elm+Tree/@43.6530556,-79.3719444,15z/data=!3m1!1e3!3m2!1s43+Elm+St,+Toronto,+ON+M5R+1A5,+Canada
TOCA	181 Wellington St W, Toronto, ON M5V 1E3, Canada	ritzcarlton.com	JJW7+55 Toronto, ON	4.3	167 reviews	https://www.google.com/maps/place/TOCA/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s181+Wellington+St+W,+Toronto,+ON+M5V+1E3,+Canada
Actinolite Restaurant	971 Ossington Ave, Toronto, ON M6H 3G7, Canada	actinoliterestaurant.com	MH9C+4V Toronto, ON	4.7	131 reviews	https://www.google.com/maps/place/Actinolite+Restaurant/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s971+Ossington+Ave,+Toronto,+ON+M6H+3G7,+Canada
STELVIO	354 Queen St W, Toronto, ON M5V 1E3, Canada	stelviotoronto.ca	JJX3+MR Toronto, ON	4.3	227 reviews	https://www.google.com/maps/place/STELVIO/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s354+Queen+St+W,+Toronto,+ON+M5V+1E3,+Canada
Nami Japanese Restaurant	55 Adelaide St E, Toronto, ON M5C 1H4, Canada	namirestaurant.ca	MJ2F+8M Toronto, ON	4.3	292 reviews	https://www.google.com/maps/place/Nami+Japanese+Restaurant/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s55+Adelaide+St+E,+Toronto,+ON+M5C+1H4,+Canada
DaiLo	503 College St, Toronto, ON M5T 1R9, Canada	dailoto.com	MH4R+83 Toronto, ON	4.6	348 reviews	https://www.google.com/maps/place/DaiLo/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s503+College+St,+Toronto,+ON+M5T+1R9,+Canada
Alo	163 Spadina Ave, Toronto, ON M5S 1A5, Canada	alorestaurant.com	JJX3+CJ Toronto, ON	4.7	553 reviews	https://www.google.com/maps/place/Alo/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s163+Spadina+Ave,+Toronto,+ON+M5S+1A5,+Canada
Scaramouche Restaurant	1 Benvenuto Pl, Toronto, ON M5V 1E3, Canada	scaramouchereastaurant.com	MHJX+HV Toronto, ON	4.7	519 reviews	https://www.google.com/maps/place/Scaramouche+Restaurant/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s1+Benvenuto+Pl,+Toronto,+ON+M5V+1E3,+Canada
Woods Restaurant	45 Colborne St, Toronto, ON M5E 1A5, Canada	woodsrestaurant.ca	JJXF+HX Toronto, ON	4.6	159 reviews	https://www.google.com/maps/place/Woods+Restaurant/@43.64749,-79.4186653,14z/data=!3m1!1e3!3m2!1s45+Colborne+St,+Toronto,+ON+M5E+1A5,+Canada

Botsol: Google Maps Crawler



FREE VERSION

The free version scrapes less data fields for more basic information than the full version, but there's no time limitation and you can use it as long as needed.

It will extract the following fields:

- Business Name
- Full Address
- Website
- Plus Code
- Rating
- Review Count
- URL

[Download Free Version](#)



FULL VERSION

Price of the full version is \$50 , It will extract the following fields.

- Business Name
- Full Address
- Street Address
- City
- State
- Zip
- Plus Code
- Website
- Phone
- Email
- Latitude, Longitude
- Category
- Hours
- Has Virtual Tour ?
- Is Claimed
- Rating
- Review Count
- Amenities
- Number of Photos
- Image URL
- URL

With ability to automatically make multiple searches

[Buy Full Version for \\$50](#)
(Single user license valid for 1 year)

Botsol: Google Review Crawler



HOW IT WORKS

Here is a brief description of how to use this app to get all reviews data:

- Launch the app.
- Search for businesses on Google Maps (from the Chrome window opened by the app).
- Click Start Button
- Wait while the crawler scrapes all reviews for each business shown in the search results.
- The app will open each business in the search results and then open the reviews; sort them by 'Newest' and it will keep scrolling down until all reviews are shown. When all reviews are shown, the app will extract and add them to the grid, which can later be exported once the bot stops.
- When the bot is done working, you can export the scraped data to a CSV or Excel file. You can also manually stop the bot if needed.



WHAT IT EXTRACTS?

Business Fields

- Business Name
- Rating
- Latitude
- Longitude
- URL

Reviews Fields

- Reviewer Name
- Date
- Publish date
- Stars
- Review Text

Parsehub: link utili

<https://help.parsehub.com/hc/en-us>

<https://help.parsehub.com/hc/en-us/articles/218181287-Create-your-first-project>

<https://help.parsehub.com/hc/en-us/articles/218185147-Select-extract-multiple-similar-elements>

<https://help.parsehub.com/hc/en-us/articles/218185147-Select-extract-multiple-similar-elements>

Parsehub

The screenshot displays the Parsehub user interface. On the left is a dark sidebar menu with the following items: Projects, Runs, My Account, Integrations, Plans & Billing, Tutorials, Documentation, API, Contact, and Log Out. The main content area is divided into three sections:

- Recent projects:** Features an illustration of a wrench, screwdriver, and hammer, a '+ New Project' button, and a list of three projects: 'tripadvisor.it Project' (two instances) and 'parsehub.com Project'. Each project has edit and share icons. A '+ See more' link is at the bottom.
- Interactive Tutorials:** A list of four tutorials: 'Learn the Basics (8 min)' (checked), 'Select and Download Data (3 min)', 'Group data with Relative Select (3 min)', and 'Click and Navigate to Links (4 min)'.
- Written Tutorials:** A list of three tutorials: 'Parsehub 101', 'Pagination ('next' page buttons)', and 'Scrape product details'.

A chat icon is visible in the bottom right corner of the interface.

main_template

- Select page +
- Select selection1 (1) [trash icon]
- Click each selection1 item
- Select selection2 [list icon] +
- Extract name
- Click each selection2 item
- and go to results_template1

Get Data

results_te...

Selection Node: Edit

All elements with class attractions-attraction-overview-main-TopPOIs_see_more--2Vsb-

Wait up to 60 seconds for elements to appear.

Quali sono le date del viaggio? Data di inizio Data

76 € per adulto 130 € per adulto 250 € per adulto

Le principali attrazioni a Iglesias



SITI D'INTERESSE
Porto Flavia
672 recensioni

Mappa



DIV
SITI D'INTERESSE

Live preview is disabled when working on interactive templates. Use Test Run instead.

main_template

results_te...

- Select page (1) +
- Select & Extract selection3 +
- Select selection4 +
- Extract name
- Relative selection5 +
- Relative selection6 +
- Select selection7
- Click each selection7 item
- and go to results_template1

Get Data

Selection Node:
1st body

Quali sono le date del viaggio?

Data di inizio

Data di fine

Select Mode

Le principali attrazioni a Iglesias

Ordina per: Preferiti dai viaggiatori



SITI D'INTERESSE

Porto Flavia

672 recensioni

Vedi 5 esperienze



SITI D'INTERESSE

Il Belvedere di Nebida

407 recensioni

Mappa

Vedi 1 esperienza

CSV/Excel JSON CSV/Excel Wide (beta)

selection3

missing data

This a live preview. When you are ready to run your project, click Get Data.



Parsehub: Results

	A	B	C	D
1	<u>selection2_name</u>	<u>selection2_selection4_name</u>	<u>selection2_selection4_selection5</u>	<u>selection2_selection4_selection6</u>
2	Porto Flavia	Per conoscere la Storia della Zona Mineraria	Interessantissima visita, ben descritta dalle guide che accomp	Data dell'esperienza: ottobre 2019
3	Porto Flavia	Paesaggio mozzafiato d'altri tempi.	Il connubio, tra la spettacolarità della visione strepitosa del Pan	Data dell'esperienza: agosto 2019
4	Porto Flavia	Un porto appeso nel vuoto	Non immaginavo una struttura così interessante e ben tenuta. I	Data dell'esperienza: novembre 2019
5	Porto Flavia	da vedere assolutamente	stupendo e interessante sito minerario che consente una bella	Data dell'esperienza: giugno 2019
6	Porto Flavia	Tornerò presto con altri amici	Grazie grazie a chi mi ha fatto conoscere questa meraviglia, co	Data dell'esperienza: ottobre 2019
7	Porto Flavia	Bello e interessante	Il luogo è fantastico e la storia che c'è dietro molto interessante	Data dell'esperienza: ottobre 2019
8	Porto Flavia	Non solo capolavoro di ingegneria.	Nell'attraversare la miniera di Porto Flavia non solo si è eruditi o	Data dell'esperienza: ottobre 2019
9	Porto Flavia	un'opera Di alta ingegneria mineraria sul mare	Splendida opera di alta ingegneria mineraria ideata dall'ingegner	Data dell'esperienza: ottobre 2019
10	Porto Flavia	Mare e non solo	Recentemente ho visitato la miniera di Porto Flavia. Visitando la	Data dell'esperienza: ottobre 2019
11	Porto Flavia	Capolavoro di ingegneria e rispetto del paesaggio	Emozionante percorrere la galleria scavata nella roccia, a picco	Data dell'esperienza: ottobre 2019
12	Porto Flavia	Sardegna poco conosciuta	Visitato il sito minerario con guida di porto Flavia, monumento a	Data dell'esperienza: settembre 2019
13	Porto Flavia	Una giornata speciale	Abbiamo visitato la miniera in una splendida mattina di settemb	Data dell'esperienza: settembre 2019
14	Porto Flavia	L'altra Sardegna	Interessantissima la visita al porto minerario Porto Flavia, nell'o	Data dell'esperienza: settembre 2019
15	Porto Flavia	Non ci hanno lasciato fare il tour	Buon giorno, ci dispiace che non abbia potuto fare la visita, fors	Data dell'esperienza: giugno 2019
16	Porto Flavia	Ammirazione	Siamo arrivati alla miniera oggi alle ore 12:05, il caldo era impeg	Data dell'esperienza: settembre 2019
17	Porto Flavia	Molto interessante	Nella visita si ripercorre il modo di trasporto, immagazzinament	Data dell'esperienza: settembre 2019
18	Porto Flavia	Un viaggio nel passato...e la meraviglia del presente	A porto Flavia ci vai così, per curiosità o per vedere da vicino il	Data dell'esperienza: settembre 2019
19	Porto Flavia	Visita a porto Flavia	La visita al porto di Torre Flavia e' molto interessante, la guida o	Data dell'esperienza: settembre 2019
20	Porto Flavia	Bella esperienza	Panorama tutto attorno stupendo, visita a Porto Flavia resa mo	Data dell'esperienza: settembre 2019
21	Porto Flavia	Meraviglia...	Una visita da non perdere, sia per la bellezza del posto, sia per	Data dell'esperienza: settembre 2019
22	Porto Flavia	Posto incantato	Bellissimo luogo, all'arrivo si può vedere la bellissima montagn	Data dell'esperienza: settembre 2019
23	Porto Flavia	Un percorso nel tempo	Bellissima esperienza con gli amici, tutti impegnati in una pass	Data dell'esperienza: agosto 2019
24	Porto Flavia	<u>Mr.</u>	Luogo stupendo, incantevole. Da fiaba, unitamente all'opera rea	Data dell'esperienza: settembre 2019
25	Porto Flavia	Galleria con vista mare	Interessante la visita alla miniera; dettagliate ed esaustive le sp	Data dell'esperienza: agosto 2019
26	Porto Flavia	Unico	L'insieme di questo luogo è davvero suggestivo: l'ingegno della	Data dell'esperienza: settembre 2019
27	Porto Flavia	Splendida vista, lente manutenzione, indicazioni fantast	Luogo davvero affascinante. Una struttura speciale, che ha sicu	Data dell'esperienza: agosto 2019

Octoparse

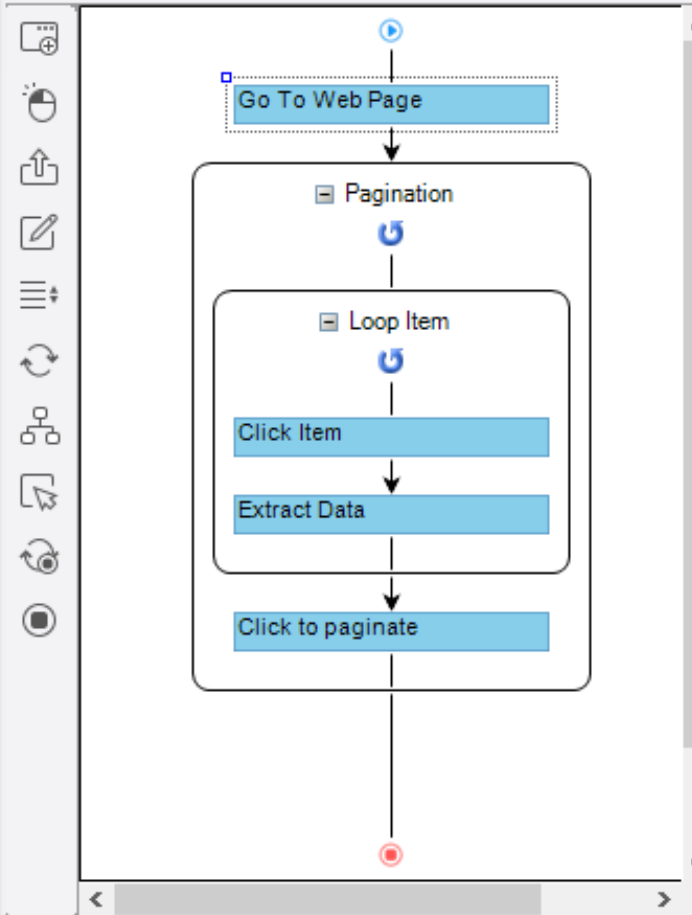
The screenshot displays the Octoparse web application interface. On the left is a vertical navigation menu with icons and labels for: Dashboard, Tools, Tutorials, Data Service, Contact Support, and About Us. At the top of the main content area is a 'Home' header. Below this, there are two primary feature cards: 'Task Templates' and 'Advanced Mode'. The 'Task Templates' card includes a circular icon of a document with a checkmark, the title 'Task Templates', the text 'Use built-in task templates to quickly get started with Octoparse', and a '+ Task' button. The 'Advanced Mode' card includes a circular icon of a document with a gear, the title 'Advanced Mode', the text 'Extract from any complex websites easily with highly flexible configuration', and a '+ Task' button with a dropdown arrow. Below these cards is a 'Tutorials' section with the text 'More Tutorials' on the right. This section contains four tutorial cards: 1) 'What's new in Octoparse 7.X?' with a cloud icon containing logos for YP, ebay, yelp, and amazon; 2) 'Capture a list of items' with an icon of overlapping document pages; 3) 'Capture data from each item page' with an icon of a document and a table; 4) 'Capture data from multiple pages' with a diagram showing a 'Pagination' box and a 'Click to paginate' button with a downward arrow. At the bottom left, there is a promotional banner for upgrading to a premium plan, with the text 'Upgrade to a premium plan to nlock more advanced features!' (note the typo 'nlock'), an 'Upgrade Now' button, and 'Current Subscription: Free'.

Octoparse

https://www.youtube.com/watch?v=j_JWaMnsXWQ&feature=youtu.be

https://www.yellowpages.com/search?search_terms=auto+repair&geo_location_terms=san+francisco

Save Start extraction Setting Best30AutoRepairinSa Workflow



Action Caption: Timeout:

Page Url: Extract

Basic options

Wait before execution: second(s) or wait until element is found

Auto Retry: Retry when page fails to load

Customize Cookie: Use specified Cookie

Use loop URLs: Open the URL in Loop Item

Advanced options

Block Pop-up: Block Pop-up Window (Possible Ads)

Scroll Down: Scroll down to bottom of the page when finished loading

Clear cache: Clear cache before loading a web page

- Dashboard
- Tools
- Tutorials
- Data Service
- Contact Support
- About Us

Upgrade to a premium plan to lock more advanced features!

[Upgrade Now](#)

The Real Yellow Pages[®]

Edit task

Extraction settings

Opening https://www.yellowpages.com/san-francisco-ca/mip/selecta-auto-body-481046704

yp The Real Yellow Pages®

Browse ▾

auto repair

San Francisco, CA

Find

Log In · Sign Up

Home > CA > San Francisco > Selecta Auto Body

Selecta Auto Body

Data extracted

	Title	Address	Phone	Website_link	Field4
1	Richardson Automotive ...	2 Richardson Ave, San ...	(415) 929-1210		
2	Don's Auto Service	899 San Jose Ave, San ...	(415) 282-0800	Visit Website	https://i2.ypcdn.com/blo...
3	All-Pro Mechanic	1357 Harrison St, San ...	(415) 487-9014	Visit Website	https://i2.ypcdn.com/blo...
4	19th Ave Chevron Servi...	1401 19th Ave, San Fra...	(415) 681-3860	Visit Website	https://i2.ypcdn.com/blo...
5	Midas Auto Repair San ...	165 Van Ness Ave, San ...	(415) 626-8384	Visit Website	
6	Sunset Service Super L...	2095 19th Ave, San Fra...	(415) 731-7211		https://i1.ypcdn.com/blo...
7	House Of Brakes Inc.	3195 24th St, San Fran...	(415) 285-0595	Visit Website	https://i2.ypcdn.com/blo...
8	Phaedrus BMW	1641 Jackson St, San F...	(415) 539-0450	Visit Website	https://i3.ypcdn.com/blo...
9	Golden Gate Tow	355 Barneveld Ave, San...	(415) 826-8866	Visit Website	https://i4.ypcdn.com/blo...
10	Ed's Autohaus	980 Harrison St, San Fr...	(415) 222-6900	Visit Website	https://i1.ypcdn.com/blo...

< 1 2

>

Data extracted: 20 lines Total time spent: 2min 51sec Average speed: 7 lines of data/min

Stop