SISTEMI INFORMATIVI E DBMS

CORSO DI LAUREA MAGISTRALE IN
MANAGEMENT E MONITORAGGIO DEL TURISMO SOSTENIBILE

ESERCITAZIONE 5

Tool di Web Scraping

Ing. Gavina Baralla

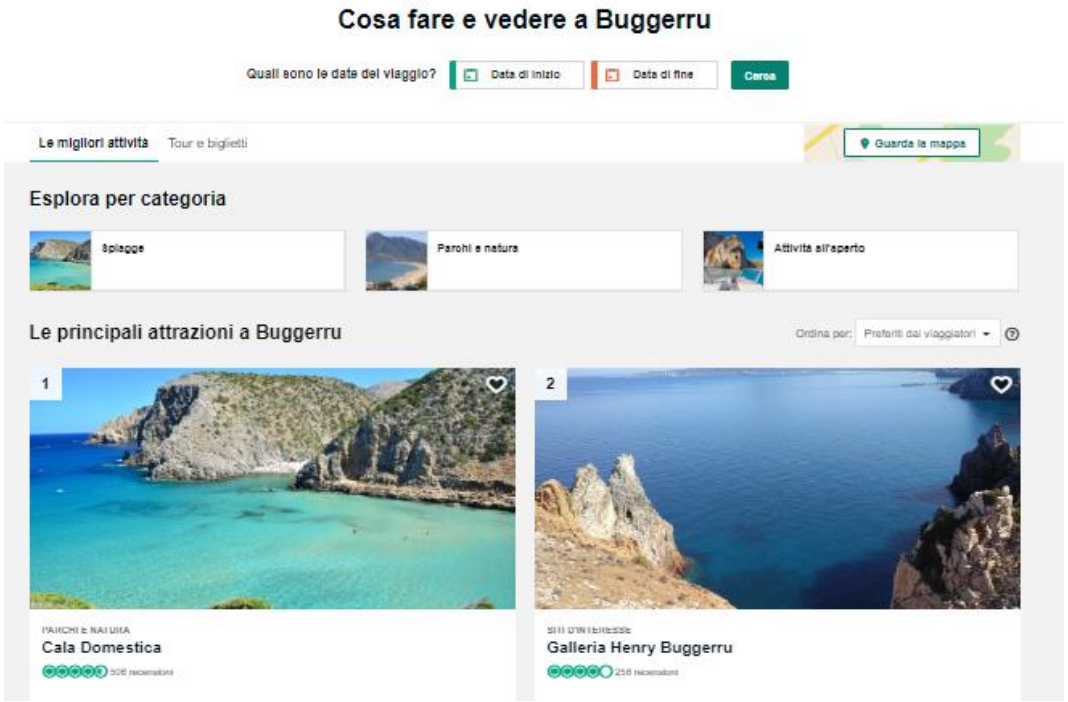
Web scraping

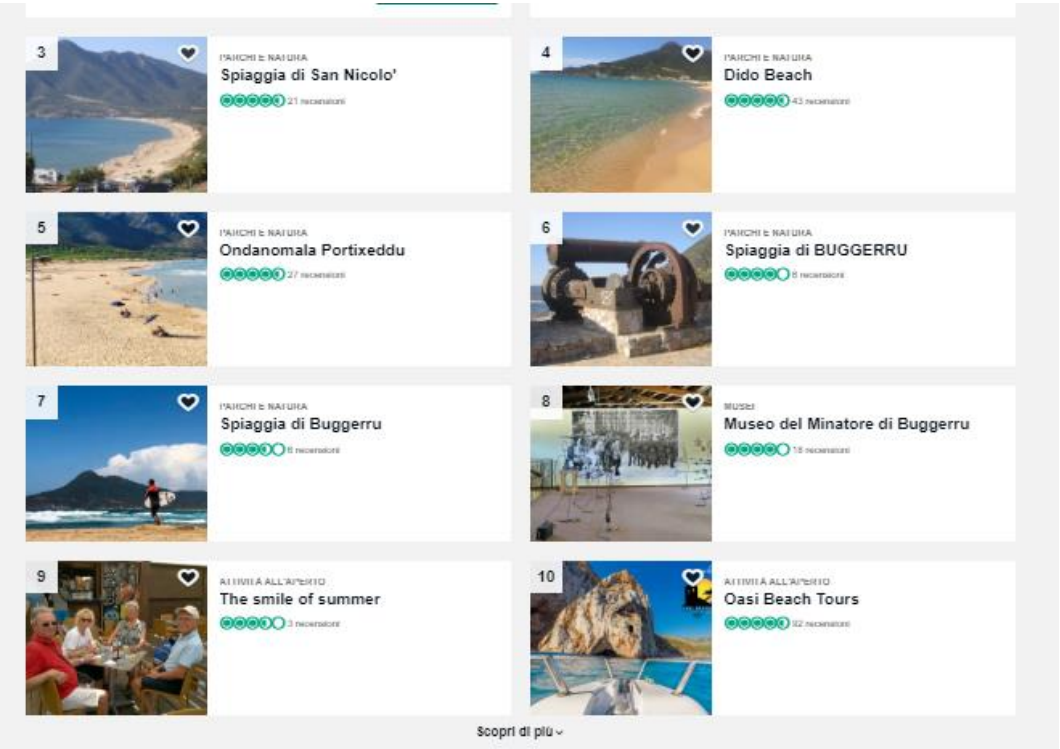
Tecnica informatica che consente l'estrazione dati da un sito web.

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





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-E

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"})
        print(containersT)
        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 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"})
        print(containersT)
        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")
```

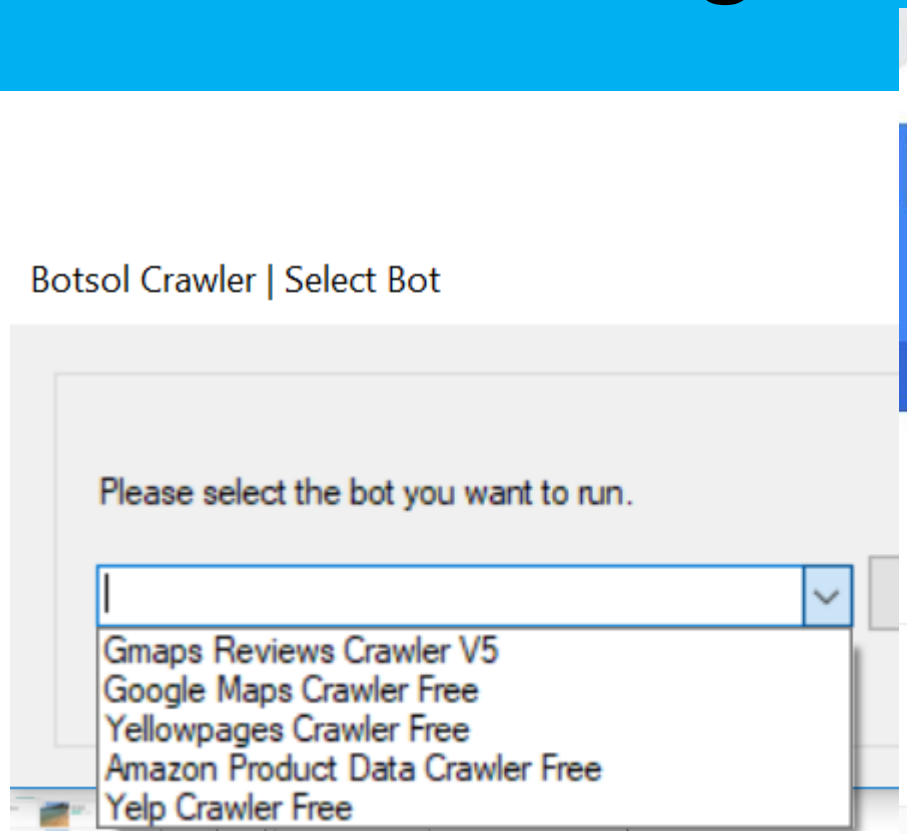
Tool di Web Scraping

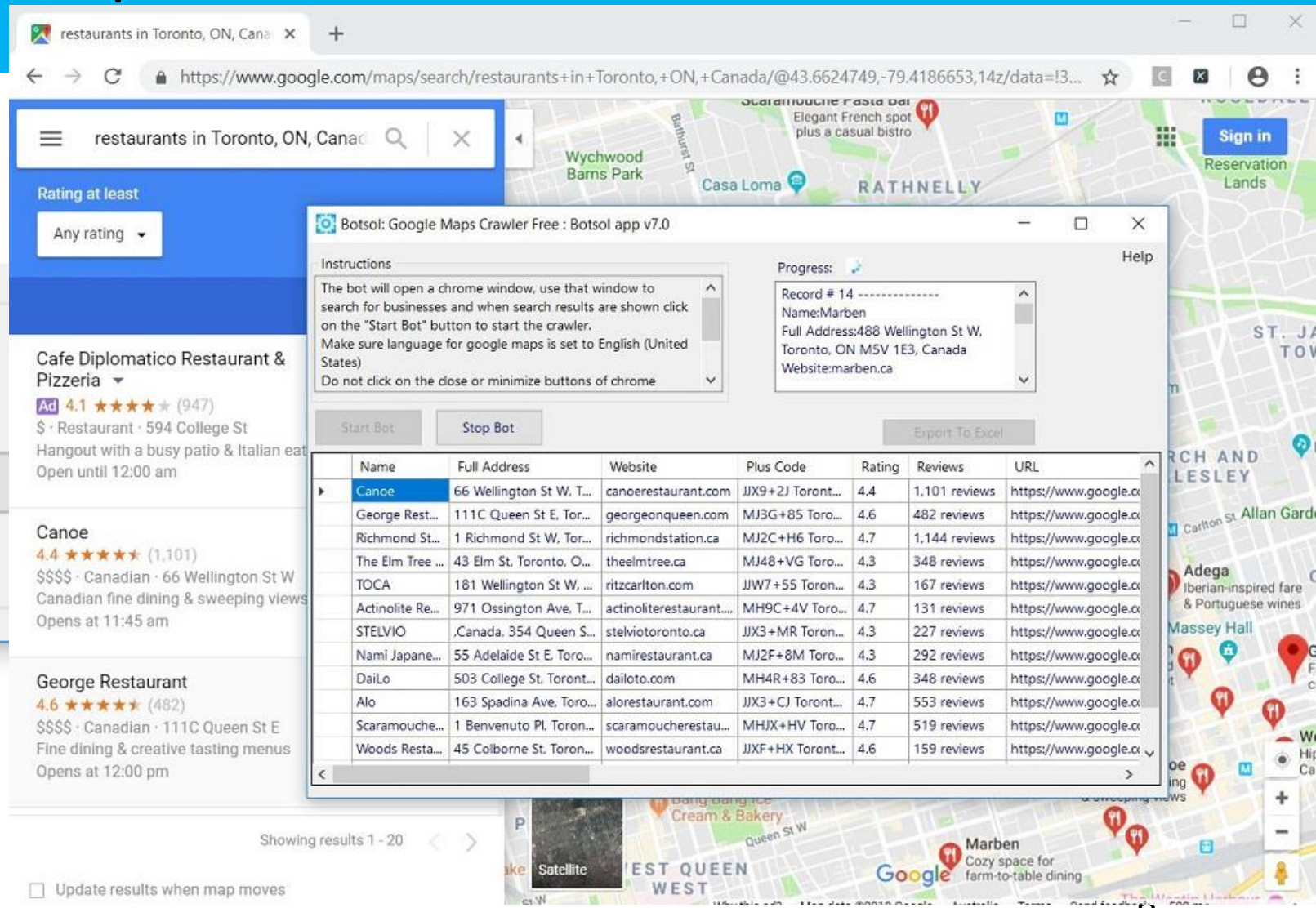
http://www.botsol.com/

https://www.octoparse.com/

https://www.parsehub.com/

Botsol: Google Maps Crawler





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

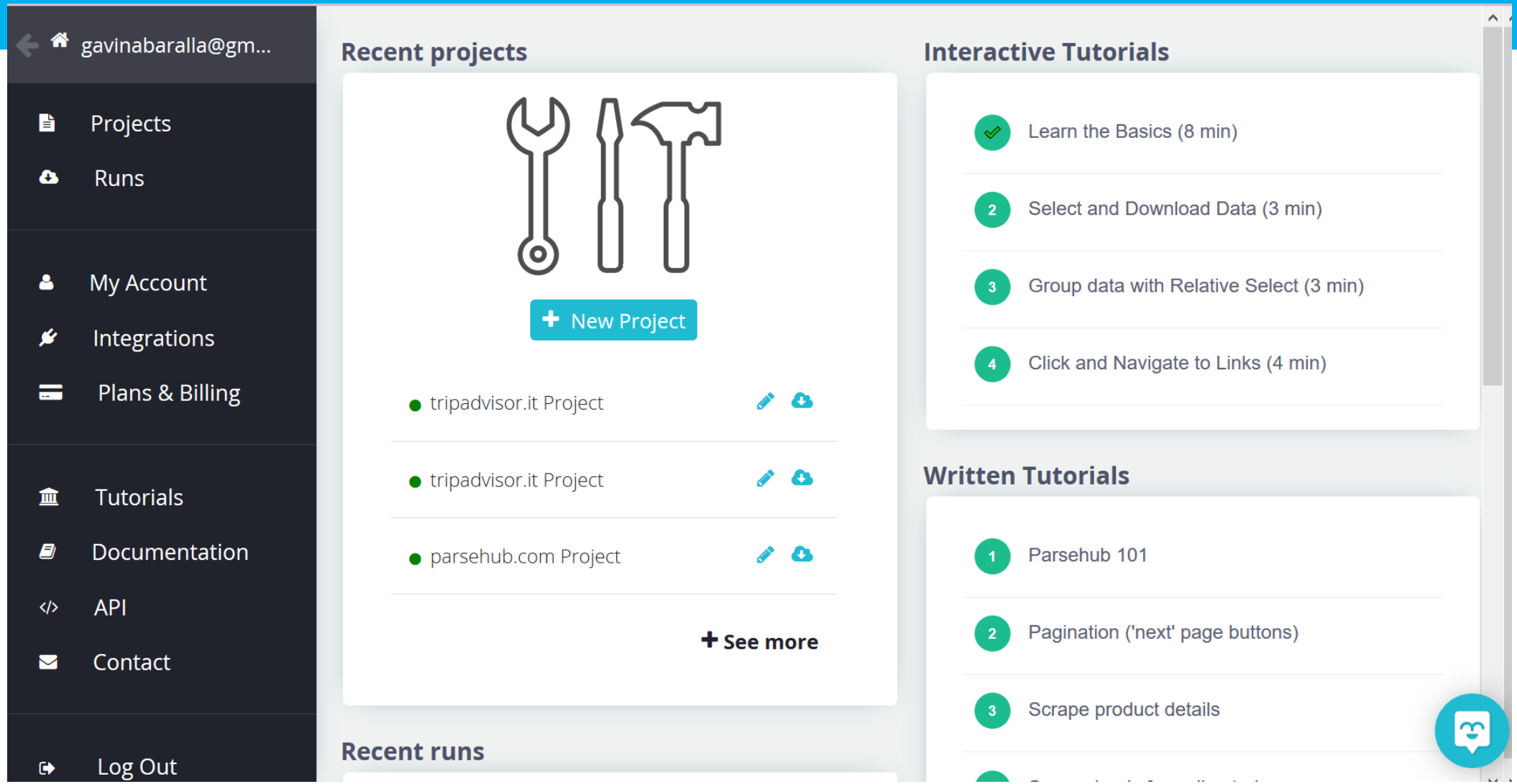
https://help.parsehub.com/hc/en-us

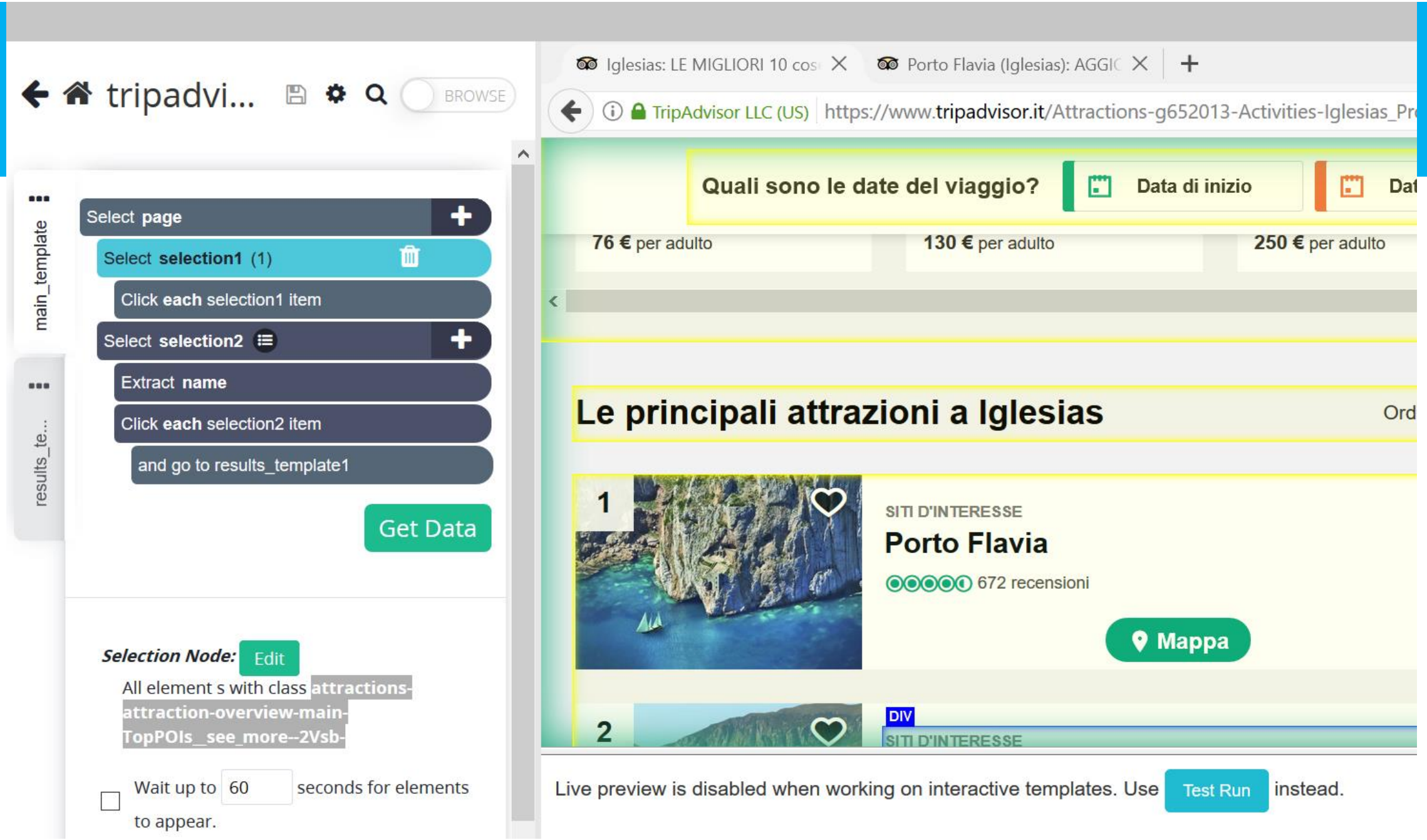
https://help.parsehub.com/hc/en-us/articles/218181287-Create-your-first-project

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


gavinabaralla@gm...
Projects
Runs
My Account
Integrations
Plans & Billing
Tutorials
Documentation
API
Contact
Log Out
Recent projects
New Project
tripadvisor.it Project
tripadvisor.it Project
parsehub.com Project
See more
Recent runs
Interactive Tutorials
Learn the Basics (8 min)
Select and Download Data (3 min)
Group data with Relative Select (3 min)
Click and Navigate to Links (4 min)
Written Tutorials
Parsehub 101
Pagination ('next' page buttons)
Scrape product details


tripadvi...
BROWSE
main_template
results_te...
Select page
Select selection1 (1)
Click each selection1 item
Select selection2
Extract name
Click each selection2 item
and go to results_template1
Get Data
Selection Node:
Edit
All element s with class attractions-attraction-overview-main-TopPOIs__see_more--2Vsb-
Wait up to 60 seconds for elements to appear.
Iglesias: LE MIGLIORI 10 cos
Porto Flavia (Iglesias): AGGIC
TripAdvisor LLC (US)
https://www.tripadvisor.it/Attractions-g652013-Activities-Iglesias_Pr
Quali sono le date del viaggio?
Data di inizio
76 € per adulto
130 € per adulto
250 € per adulto
Le principali attrazioni a Iglesias
1
SITI D'INTERESSE
Porto Flavia
672 recensioni
Mappa
2
DIV
SITI D'INTERESSE
Live preview is disabled when working on interactive templates. Use
Test Run
instead.

tripadvi...

BROWSE

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**

Iglesias: LE MIGLIORI 10 cos

TripAdvisor LLC (US) https://www.tripadvisor.it/Attractions-g652013-Activities-Iglesias_Province_of_Carbonia_Iglesi

Quali sono le date del viaggio? Data di inizio Data di fine Select Mode

Le principali attrazioni a Iglesias

Ordina per: Preferiti dai viaggiatori

1

SITI D'INTERESSE

Porto Flavia

672 recensioni

Vedi 5 esperienze

2

SITI D'INTERESSE

Il Belvedere di Nebida

407 recensioni Mappa

Vedi 1 esperienza

CSV/Excel JSON CSV/Excel Wide (beta)

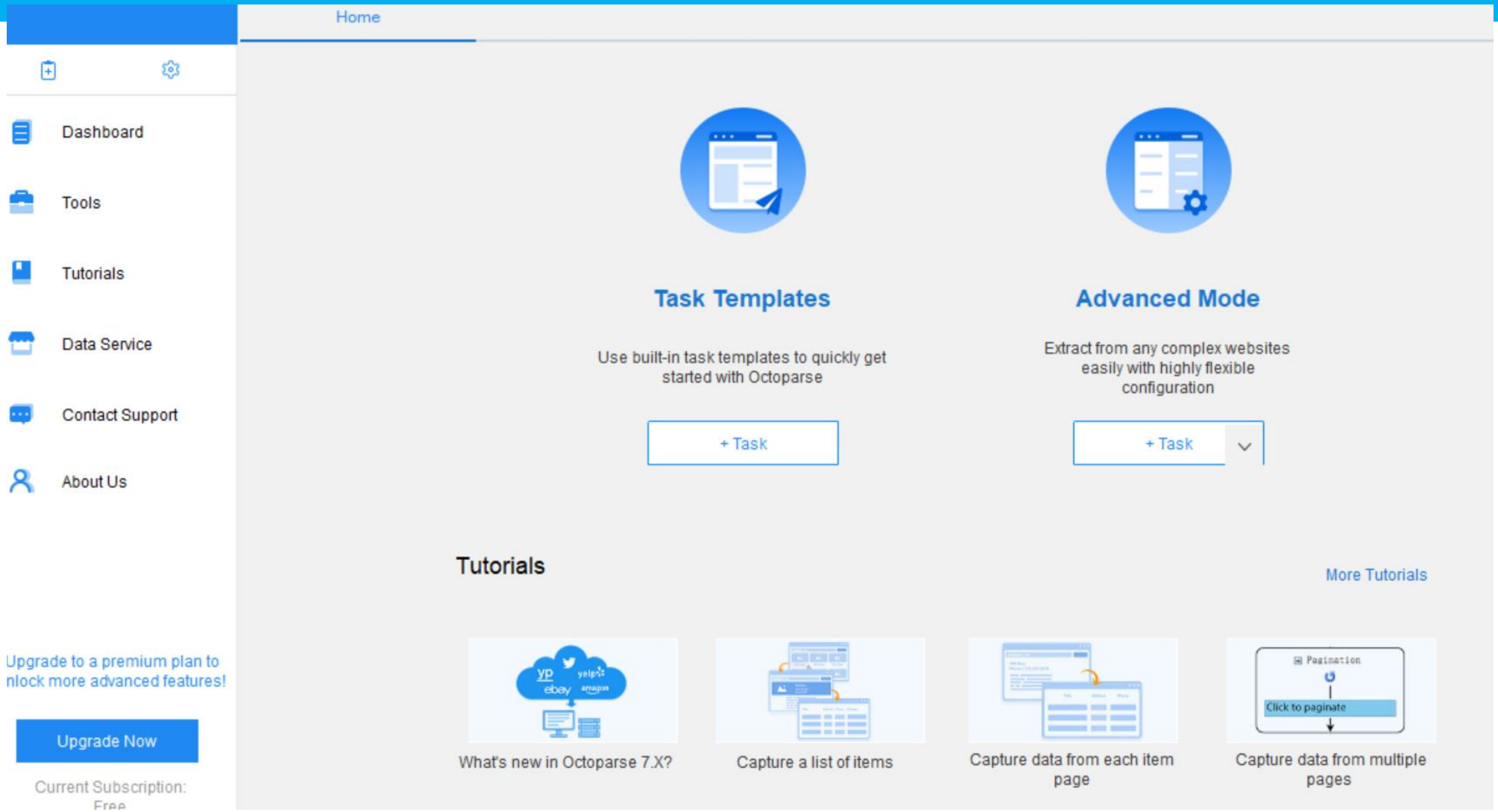
selection3 **missing data**

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

Parsehub: Results

	A	B	C	D
1	selection2_name	selection2_selection4_name	selection2_selection4_selection5	selection2_selection4_selection6
2	Porto Flavia	Per conoscere la Storia della Zona Mineraria	Interessantissima visita, ben descritta dalle guide che accompa	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 c	Data dell'esperienza: ottobre 2019
9	Porto Flavia	un'opera Di alta ingegneria mineraria sul mare	Spledita opera di alta ingegneria mineraria ideata dall'ingegner V	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'lo	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 c	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 montagna	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	Mr.	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 spi	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, lenta manutenzione, indicazioni fant	Luogo davvero affascinante. Una struttura geniale, che ha sicu	Data dell'esperienza: agosto 2019

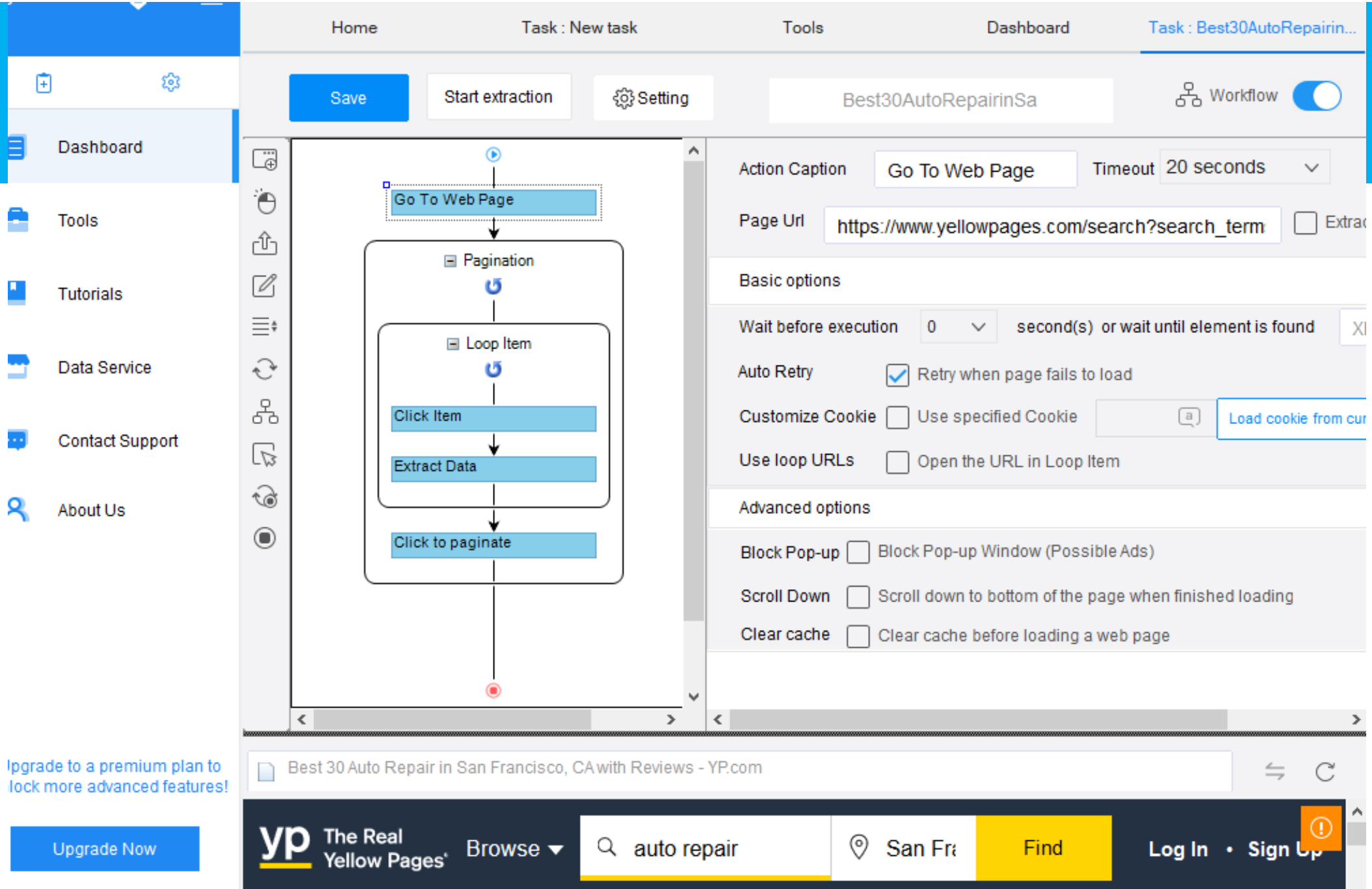
Octoparse

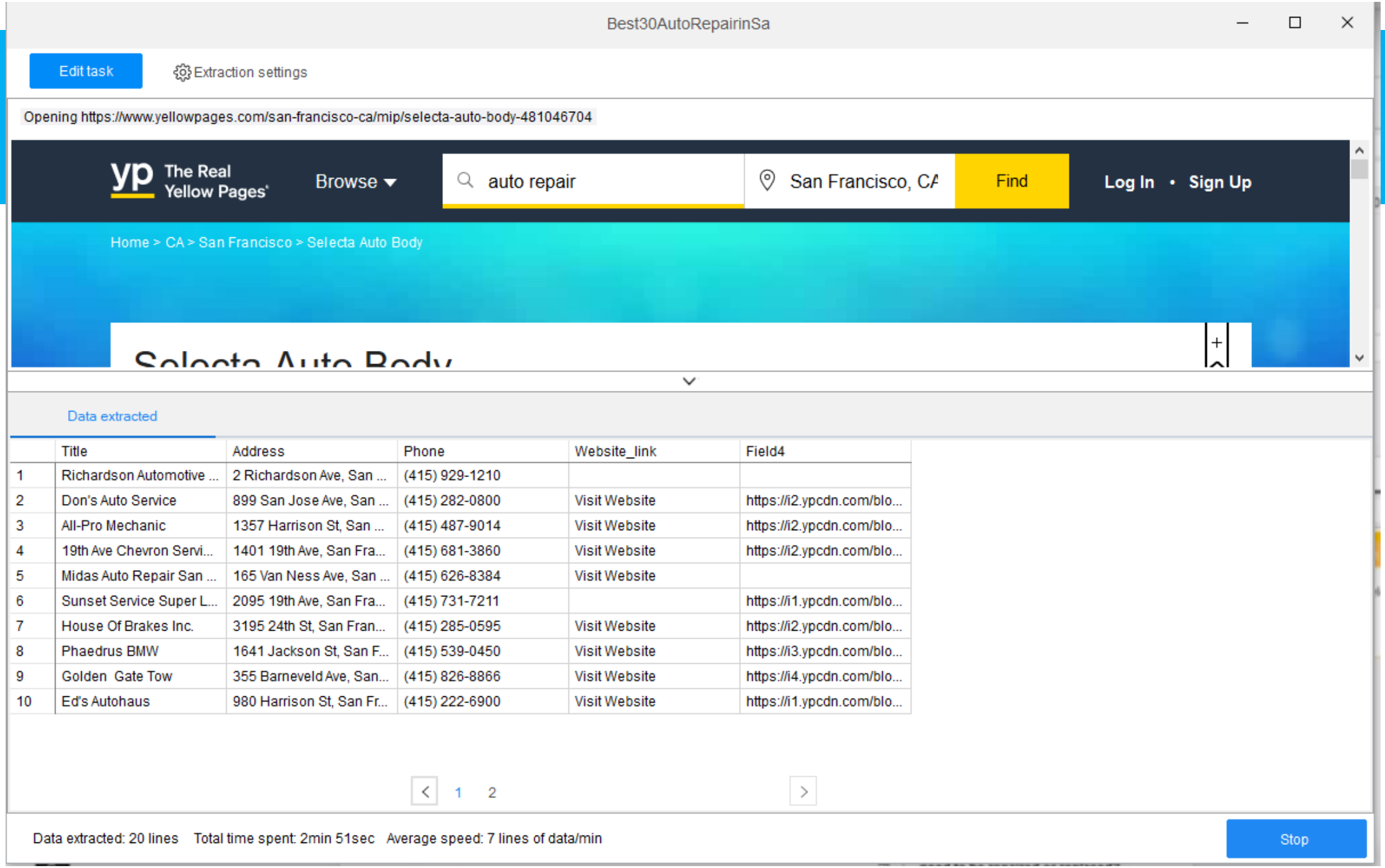

Home
Dashboard
Tools
Tutorials
Data Service
Contact Support
About Us
Upgrade to a premium plan to nlock more advanced features!
Upgrade Now
Current Subscription:
Free
Task Templates
Use built-in task templates to quickly get started with Octoparse
+ Task
Advanced Mode
Extract from any complex websites easily with highly flexible configuration
+ Task
Tutorials
More Tutorials
What's new in Octoparse 7.X?
Capture a list of items
Capture data from each item page
Pagination
Click to paginate
Capture data from multiple pages

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


Home
Task : New task
Tools
Dashboard
Task : Best30AutoRepairin...
Save
Start extraction
Setting
Best30AutoRepairinSa
Workflow
Dashboard
Tools
Tutorials
Data Service
Contact Support
About Us
Go To Web Page
Pagination
Loop Item
Click Item
Extract Data
Click to paginate
Action Caption
Go To Web Page
Timeout
20 seconds
Page Url
https://www.yellowpages.com/search?search_term
Basic options
Wait before execution
0
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
Best 30 Auto Repair in San Francisco, CA with Reviews - YP.com
Upgrade to a premium plan to
lock more advanced features!
Upgrade Now
The Real
Yellow Pages
Browse
auto repair
San Fra
Find
Log In
Sign Up


Best30AutoRepairinSa
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