

# TRIPADVISOR SCRAPING SERVICE



**web scraping**

Rubén Grávalos

10-11-2023

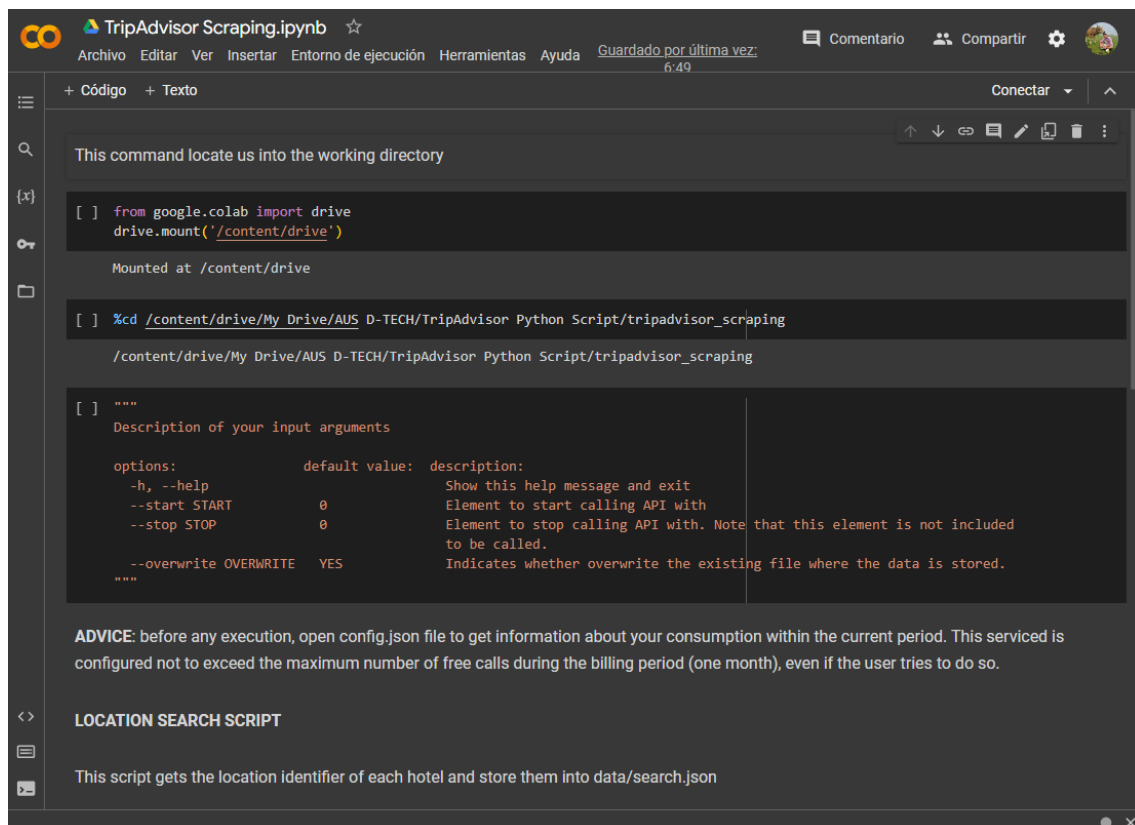
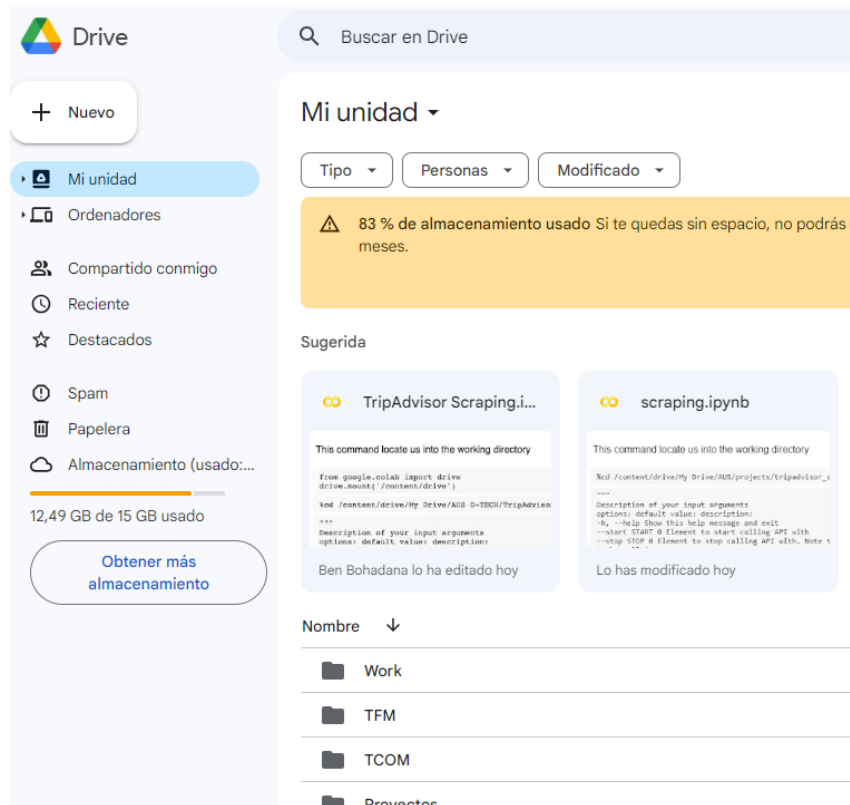
## Contenido

LOCATING THE PROJECT FILES .....	2
1- Open the Colab notebook. ....	2
2- Locate into the working directory. ....	3
(ONLY ONCE) Enter your TripAdvisor API User Key.....	4
Run the Colab Script.....	5
Locate the Colab execution environment into the project directory. ....	5
Understand the input parameters.....	5
3. Location Search Script .....	7
4. Location Details Script. ....	9
5. Location Reviews Script. ....	10
Manage de .csv data.....	11

## LOCATING THE PROJECT FILES

1- Open the Colab notebook.

1.1. Login to your Google Drive account in your browser and open your Colab notebook (i.e. “TripAdvisor Scrapping.ipynb”).



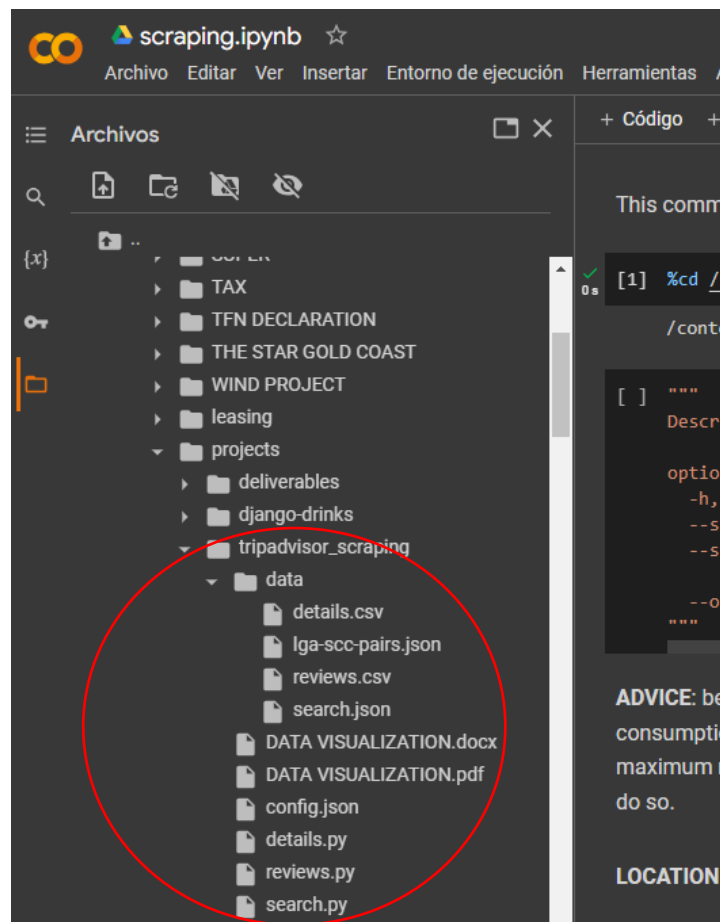
## 2- Locate into the working directory.

2.1. Run the following section.

```
from google.colab import drive
drive.mount('/content/drive')
```

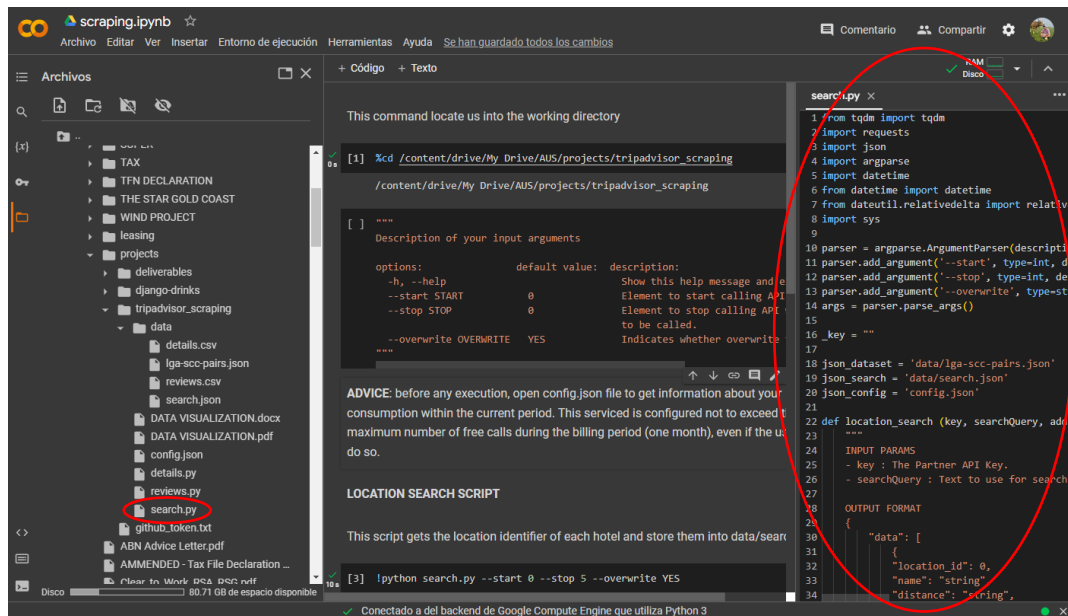
Mounted at /content/drive

2.2. Navigate to the working directory.



(ONLY ONCE) Enter your TripAdvisor API User Key.

1. Double click on “search.py” file and a text editor will be opened at the right side of the screen.



- 2- Copy your TripAdvisor User API Key from <https://www.tripadvisor.com/developers?screen=credentials> and paste it in line 16 between the commas as in the next example:

```
search.py X
14 args = parser.parse_args()
15
16 _key = "YOUR_API_KEY_HERE"
17
```

- 3- Ensure the changes are saved checking the name of the file, as in the above picture is “search.py”. This name means that the changes has been successfully saved. If the name of the file is shown as “\*search.py”, the symbol \* means that the changes hasn’t been saved yet. To save the changes, just double click on the name of the file “\*search.py”. It will turn into “search.py” meaning that the changes have been saved.

## Run the Colab Script

Locate the Colab execution environment into the project directory.

Run the following section.

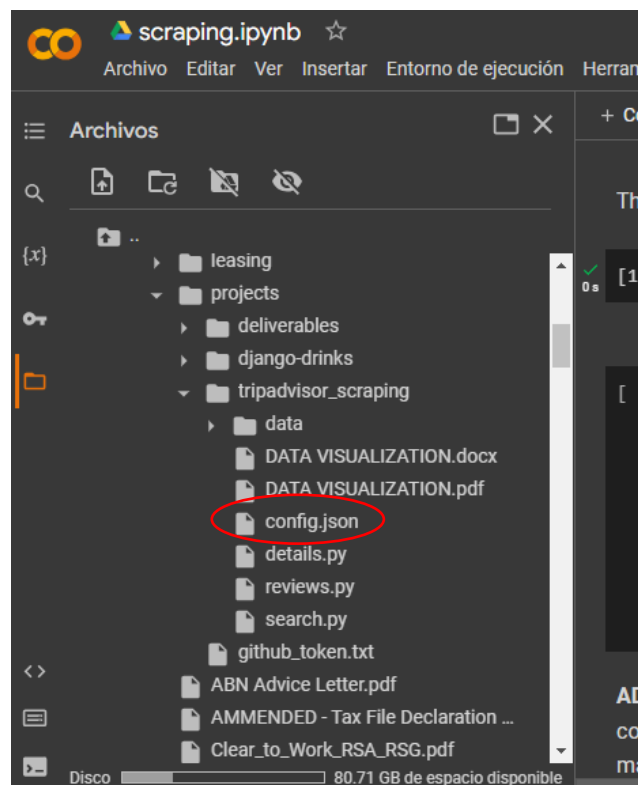
```
[ ] %cd /content/drive/My Drive/AUS D-TECH/TripAdvisor Python Script/tripadvisor_scraping  
  
/content/drive/My Drive/AUS D-TECH/TripAdvisor Python Script/tripadvisor_scraping
```

## Understand the input parameters.

```
Description of your input arguments  
  
options:          default value:  description:  
-h, --help              0          Show this help message and exit  
--start START          0          Element to start calling API with  
--stop STOP            0          Element to stop calling API with. Note that this element is not included  
                           to be called.  
--overwrite OVERWRITE YES Indicates whether overwrite the existing file where the data is stored.
```

1. Example for Location Search Script.

1.1. Open the configuration file “config.json”.



1.2. Choose the number of hotels you want to get the information from.

```
"search": {  
  "n_elements": 519,  
  "last_starting_at": 0,  
  "last_stopping_at": 5  
}
```

*n\_elements* : total number of city-suburb pairs in the dataset you are available to call. The dataset is located inside the “data” directory, in “lga-scc-pairs.json” file.

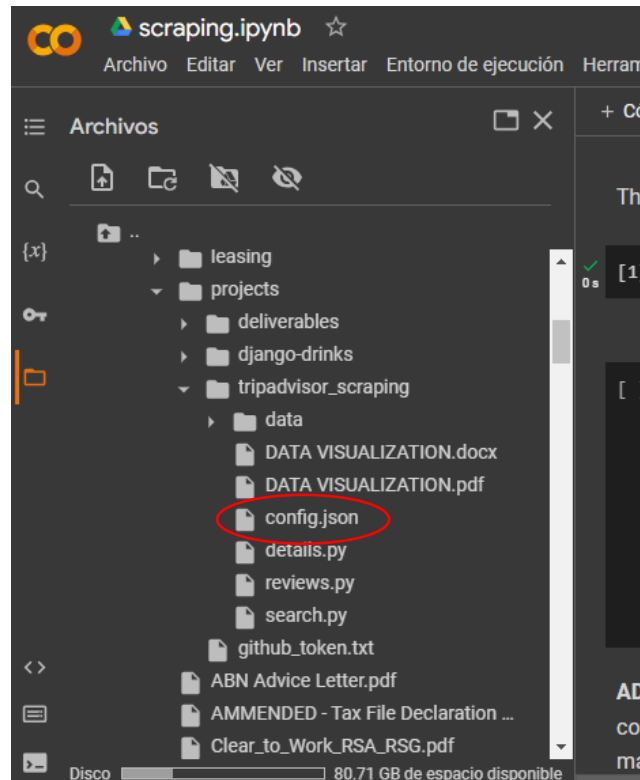
*last\_starting\_at* : The pair you started calling in the last execution.

*last\_stopping\_at* : The pair you stopped calling in the last execution.

**NOTE:** The last pair of an input range when executing the script is never called. For example, when you make a call starting at 0 and stopping at 5, you call the pairs from 0 to 4 both included. It means that in the next call, you'll need to start calling from 5. To the edge situation, when you mean to call the last pair (519), you'll need to start calling, for example, at 515, and stop calling at 520.

## 2. Example for Location Details Script.

### 2.1. Open the configuration file "config.json".



- Choose the number of hotels you want to get the information from.

```
"details": {  
  "n_elements": 24,  
  "last_starting_at": 4,  
  "last_stopping_at": 7  
}
```

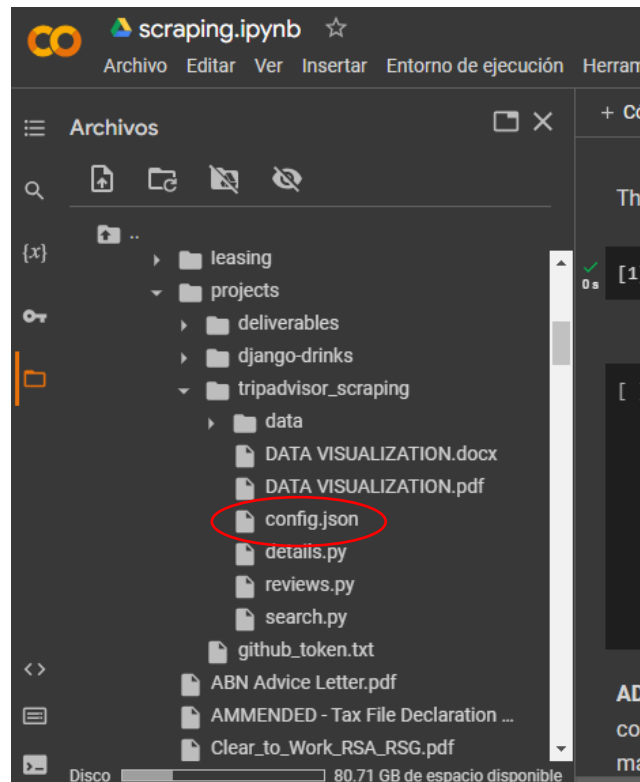
*n\_elements* : total number of location\_id saved into "search.json" you can use to make calls to get de contact details.

*last\_starting\_at* : The pair you started calling in the last execution.

*last\_stopping\_at* : The pair you stopped calling in the last execution.

**NOTE:** The last pair of an input range when executing the script is never called. For example, when you make a call starting at 0 and stopping at 5, you call the pairs from 0 to 4 both included. It means that in the next call, you'll need to start calling from 5. To the edge situation, when you mean to call the last pair (24), you'll need to start calling, for example, at 20, and stop calling at 25.

3. Example for Location Reviews Script.
- 2.2. Open the configuration file “config.json”.



- b. Choose the number of hotels you want to get the information from.

```
"reviews": {  
  "n_elements": 3,  
  "last_starting_at": 0,  
  "last_stopping_at": 0
```

*n\_elements* : total number of location\_id saved into “details.csv” you can use to make calls to get de reviews of the locations.

*last\_starting\_at* : The pair you started calling in the last execution.

*last\_stopping\_at* : The pair you stopped calling in the last execution.

**NOTE:** The last pair of an input range when executing the script is never called. For example, when you make a call starting at 0 and stopping at 5, you call the pairs from 0 to 4 both included. It means that in the next call, you’ll need to start calling from 5. To the edge situation, when you mean to call the last pair (24), you’ll need to start calling, for example, at 20, and stop calling at 25.

### 3. Location Search Script

#### 3.1. Run the section.

**NOTE:** The parameter `–overwrite` should be set to NO unless you want to delete the content of the file when running the command. You may want to set it to YES the first time you start calling from pair 0.

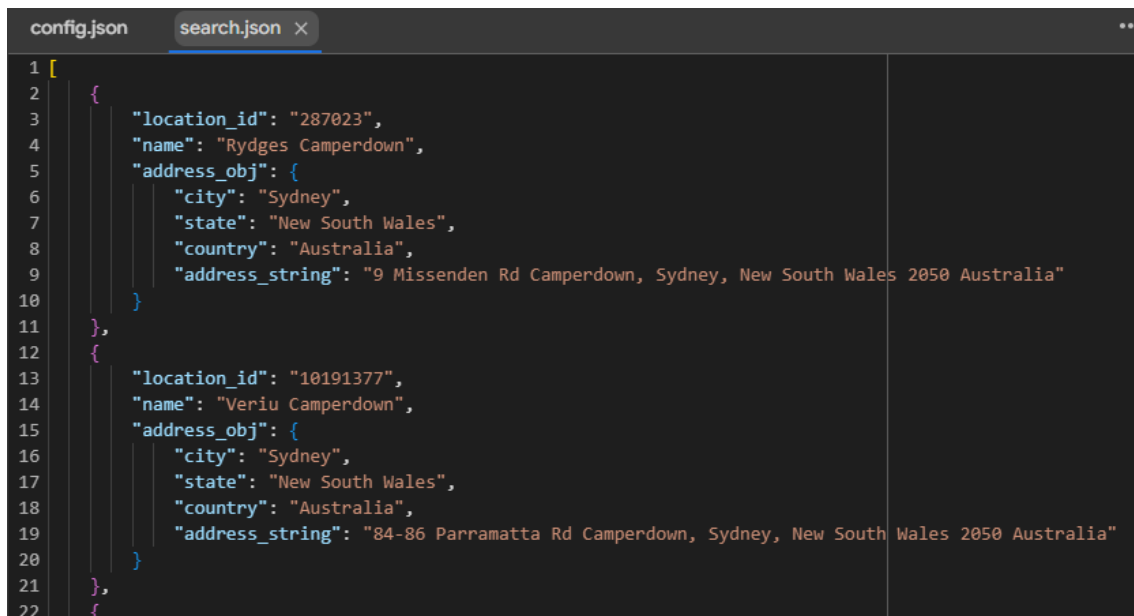


```
[3] !python search.py --start 0 --stop 5 --overwrite YES

Total number of calls for this execution: 5.
100% 5/5 [00:08<00:00, 1.68s/call]
DONE!
Number of hotels added: 24
Total number of hotels in data/search.json: 24
```

### 3.2. Watch the scraped information.

Go to the file explorer at the left of the screen and double click on “search.json” file inside the “data” directory. You will see the stored information of each hotel in .json format. You don’t have to use this file to get the contact details of the hotels, as they won’t be shown in here. Here is just stored the location\_id parameter you’ll use to make calls with that will get you the contact details of each hotel, which will be stored in a different file.



```
config.json  search.json x
1 [
2   {
3     "location_id": "287023",
4     "name": "Ryldges Camperdown",
5     "address_obj": {
6       "city": "Sydney",
7       "state": "New South Wales",
8       "country": "Australia",
9       "address_string": "9 Missenden Rd Camperdown, Sydney, New South Wales 2050 Australia"
10    }
11  },
12  {
13    "location_id": "10191377",
14    "name": "Veriu Camperdown",
15    "address_obj": {
16      "city": "Sydney",
17      "state": "New South Wales",
18      "country": "Australia",
19      "address_string": "84-86 Parramatta Rd Camperdown, Sydney, New South Wales 2050 Australia"
20    }
21  },
22  {
```

## 4. Location Details Script.

### 4.1. Run the section.

**NOTE:** The parameter `--overwrite` should be set to NO unless you want to delete the content of the file when running the command. You may want to set it to YES the first time you start calling from pair 0.

```
[11] !python details.py --start 0 --stop 5 --overwrite YES

Total number of calls for this execution: 5.
100% 5/5 [00:05<00:00, 1.05s/hotel]
DONE!
Number of hotels' details added: 5
Total number of hotels in data/details.csv: 5
```

### 4.2. Watch the scraped information.

Go to the file explorer at the left of the screen and double click on “details.csv” file inside the “data” directory. You will see the stored information of each hotel in a table format. For those cells where no data is shown, it means that there’s not such information coming from TripAdvisor.

1 to 5 of 5 entries <span>Filter</span> <span></span>											
location_id	name	web_url	street1	street2	city	state	country	postcode	email	phone	
287023	Rydges Camperdown	<a href="https://www.tripadvisor.com/Hotel_Review-g255060-d287023-Reviews-Rydges_Camperdown-Sydney_New_South_Wales.html?m=66827">https://www.tripadvisor.com/Hotel_Review-g255060-d287023-Reviews-Rydges_Camperdown-Sydney_New_South_Wales.html?m=66827</a>	9 Missenden Rd	Camperdown	Sydney	New South Wales	Australia	2050			
10191377	Veriu Camperdown	<a href="https://www.tripadvisor.com/Hotel_Review-g255060-d10191377-Reviews-Veriu_Camperdown-Sydney_New_South_Wales.html?m=66827">https://www.tripadvisor.com/Hotel_Review-g255060-d10191377-Reviews-Veriu_Camperdown-Sydney_New_South_Wales.html?m=66827</a>	84-86 Parramatta Rd	Camperdown	Sydney	New South Wales	Australia	2050			
5990761	Camperdown Hotel	<a href="https://www.tripadvisor.com/Attraction_Review-g255060-d5990761-Reviews-Camperdown_Hotel-Sydney_New_South_Wales.html?m=66827">https://www.tripadvisor.com/Attraction_Review-g255060-d5990761-Reviews-Camperdown_Hotel-Sydney_New_South_Wales.html?m=66827</a>	138 Parramatta Rd		Sydney	New South Wales	Australia	2050		9557 1615	<a href="http://www.camperdownhotel.com.au">http://www.camperdownhotel.com.au</a>
257306	Quality Apartments Camperdown	<a href="https://www.tripadvisor.com/Hotel_Review-g255060-d257306-Reviews-Quality_Apartments_Camperdown-Sydney_New_South_Wales.html?m=66827">https://www.tripadvisor.com/Hotel_Review-g255060-d257306-Reviews-Quality_Apartments_Camperdown-Sydney_New_South_Wales.html?m=66827</a>	23-33 Missenden Rd	Camperdown	Sydney	New South Wales	Australia	2050			
23387535	Camperdown	<a href="https://www.tripadvisor.com/Tourism-g23387535-Camperdown_New_South_Wales-Vacations.html?m=66827">https://www.tripadvisor.com/Tourism-g23387535-Camperdown_New_South_Wales-Vacations.html?m=66827</a>			Sydney	New South Wales	Australia				

Show 10 per page

## 5. Location Reviews Script.

### 5.1. Run the section.

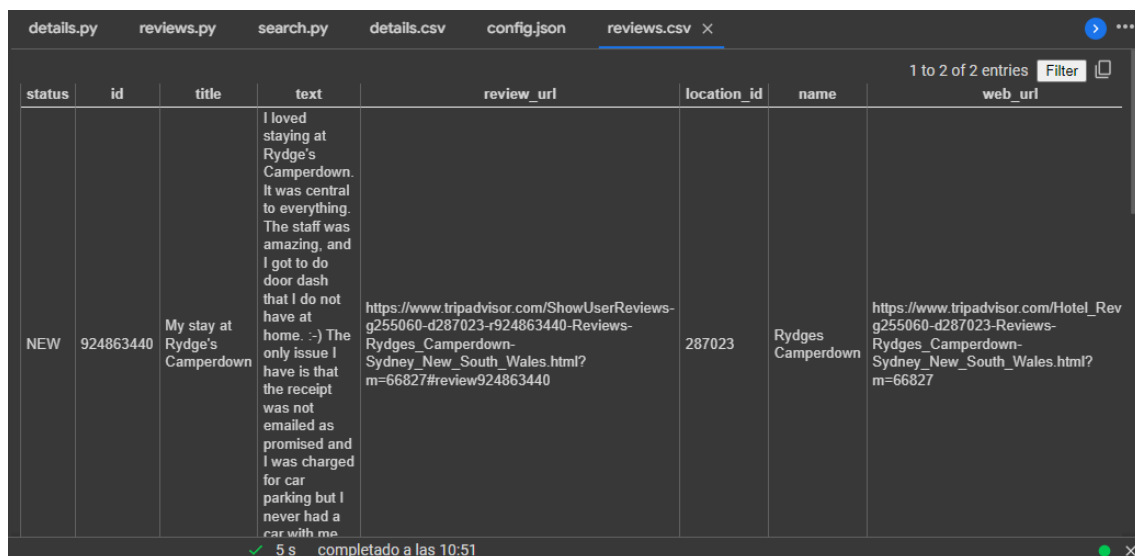
**NOTE:** The parameter `--overwrite` should be set to NO unless you want to delete the content of the file when running the command. You may want to set it to YES the first time you start calling from pair 0. If you enter NO, the OLD reviews that are scraped again will turn into OLD value for the STATUS column.

```
[15] !python reviews.py --start 0 --stop 5 --overwrite YES

Total number of calls for this execution: 5.
100% 5/5 [00:04<00:00, 1.08hotel/s]
DONE!
Number of reviews added: 2
Total number of reviews in data/reviews.csv: 2
```

### 5.2. Watch the scraped information.

Go to the file explorer at the left of the screen and double click on “reviews.csv” file inside the “data” directory. You will see the stored information of each review matching the filter criteria of finding “bug” in the title or in the text of any review, and the contact details of the hotel in a table format. If no data is shown, it means that there isn’t any review matching the filter criteria for the hotels scraped from TripAdvisor.



status	id	title	text	review_url	location_id	name	web_url
NEW	924863440	My stay at Rydges Camperdown	I loved staying at Rydges Camperdown. It was central to everything. The staff was amazing, and I got to do door dash that I do not have at home. :) The only issue I have is that the receipt was not emailed as promised and I was charged for car parking but I never had a car with me	<a href="https://www.tripadvisor.com/ShowUserReviews-g255060-d287023-r924863440-Reviews-Rydges_Camperdown-Sydney_New_South_Wales.html?m=66827#review924863440">https://www.tripadvisor.com/ShowUserReviews-g255060-d287023-r924863440-Reviews-Rydges_Camperdown-Sydney_New_South_Wales.html?m=66827#review924863440</a>	287023	Rydges Camperdown	<a href="https://www.tripadvisor.com/Hotel_Review-g255060-d287023-Reviews-Rydges_Camperdown-Sydney_New_South_Wales.html?m=66827">https://www.tripadvisor.com/Hotel_Review-g255060-d287023-Reviews-Rydges_Camperdown-Sydney_New_South_Wales.html?m=66827</a>

5 s completado a las 10:51

## Manage de .csv data

You can download the .csv files to your local computer, open Excel, click on Data, Get Data, From File, From Text/CSV, select the .csv file, and press the Load button. This way, you can save de data in a Excel file for your convenience to manipulate the details.

