

Group 1

1.SOFTWARE SET-UP

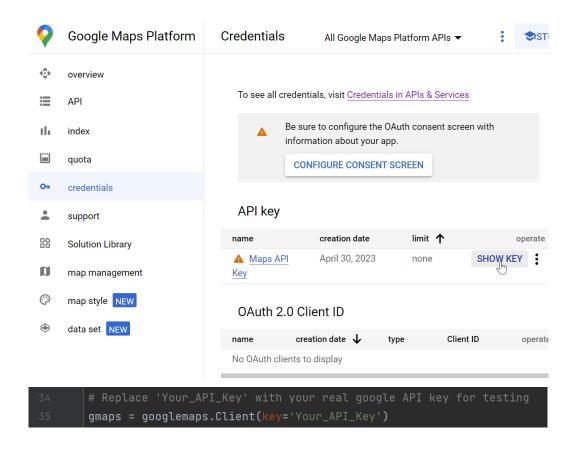
1) Installing PyCharm

Download link: https://www.jetbrains.com/pycharm/download/?section=windows.

(We didn't try PyCharm Community Edition, but it might work as well)



- 2) Create your own google API key if you don't have one.
 - a. Go to the Google Cloud Console: https://console.cloud.google.com/
 - b. Sign in with the Google account you want to associate with the API key.
 - c. Create a new project or select an existing project from the project dropdown in the top navigation bar.
 - d. In the sidebar, click on "APIs & Services" and then click on "Credentials."
 - e. On the Credentials page, click on the "Create Credentials" button and select "API Key" from the dropdown menu.
 - f. A dialog box will appear displaying your newly created API key. Take note of this key as it will be used to authenticate your requests to Google APIs.
 - g. You can further customize the API key by restricting its usage to specific IP addresses, HTTP referrers, or API restrictions. These settings provide an added layer of security and control over the key's usage.
 - h. Once you have created and configured the API key, make sure to securely store it. Do not expose your API key publicly, such as in public repositories or client-side code, as it can lead to unauthorized usage and potential security issues.
 - For more detailed instructions and information, you can refer to the official Google Cloud documentation on creating an API key: https://cloud.google.com/docs/authentication/getting-started#creating-the-api-key
 - a. Replace 'Your_API_Key' with your real google API key in app.py line 35.



- 3) Download Firefox browser and geckodriver that matches your computer.
 - b. Firefox Brower download page: https://www.mozilla.org/en-CA/firefox/new/
 - c. geckodriver download page: https://github.com/mozilla/geckodriver/releases.
 Please choose the one that matches your computer, i.e. if you are using windows 64-bit operating system, you should choose win64.zip.

♥geckodriver-v0.33.0-linux-aarch64.tar.gz	2.93 MB	Apr
geckodriver-v0.33.0-linux-aarch64.tar.gz.asc	591 Bytes	Apr
♦ geckodriver-v0.33.0-linux32.tar.gz	3.02 MB	Apr
geckodriver-v0.33.0-linux32.tar.gz.asc	833 Bytes	Арі
ੑੑੑੑੑੑ ਊgeckodriver-v0.33.0-linux64.tar.gz	2.93 MB	Арі
ੑੑੑੑੑੑੑ ਊgeckodriver-v0.33.0-linux64.tar.gz.asc	833 Bytes	Арі
♦ geckodriver-v0.33.0-macos-aarch64.tar.gz	1.85 MB	Арі
♦ geckodriver-v0.33.0-macos.tar.gz	2.05 MB	Арі
❤geckodriver-v0.33.0-win-aarch64.zip	1.47 MB	Арі
ੑੑੑੑੑੑੑ o geckodriver-v0.33.0-win32.zip	1.54 MB	Арі
ੑੑੑੑੑ ਊgeckodriver-v0.33.0-win64.zip	1.59 MB	Арі
Source code (zip)		Apr

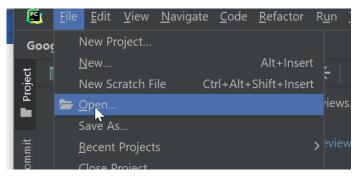
- d. You could use <u>test_selenium.py</u> to check if you installed a matched version. If so when you run it, you will see it open a Firefox browser window on google (it will be closed quickly which is expected).
- e. Also set path to **geckodriver** as per your configuration, change it to your path where you save the **geckodriver** accordingly in app.py line 257.

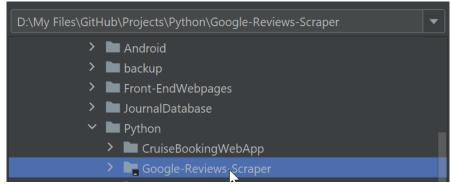
Set path to geckodriver as per your configuration, change it to your p<u>at</u>

webdriver_service = Service(r'D:\My Files\download\geckodriver.exe')

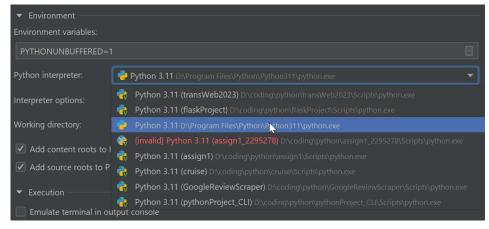
2. SCRAPING STEPS

1) Open this scraper with PyCharm

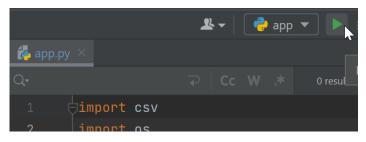




2) Set the configuration and install all necessary packages.



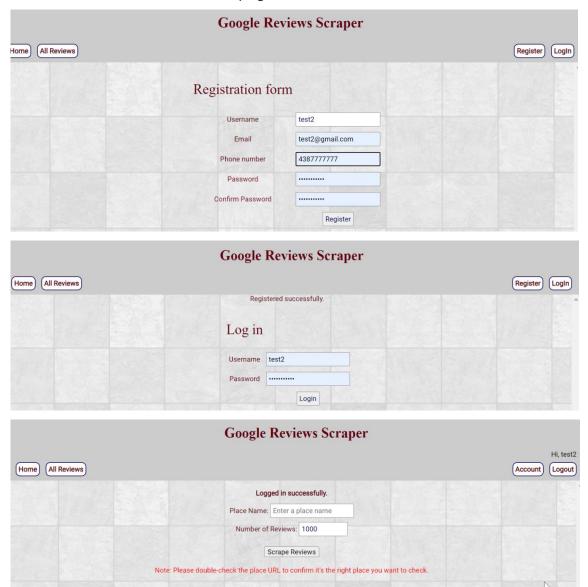
3) Open app.py and click the "run" button on the right top



4) Click the development server link in app console (http://127.0.0.1:5000).

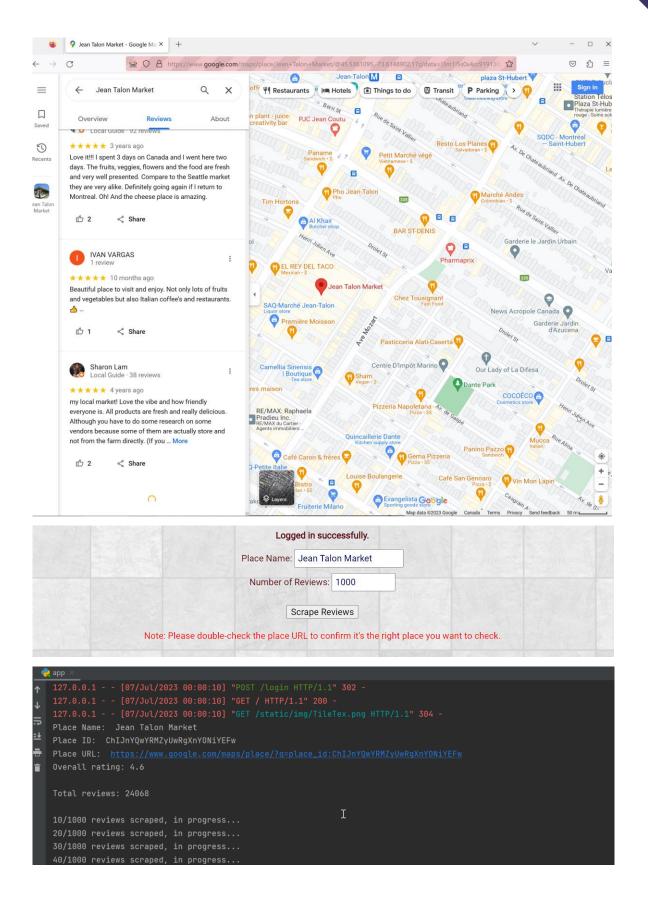


5) Register and login. You could edit your account information if you want with the "Account" button and "/account" page.



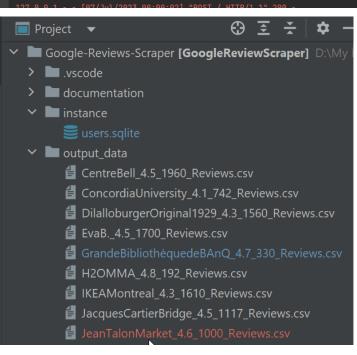


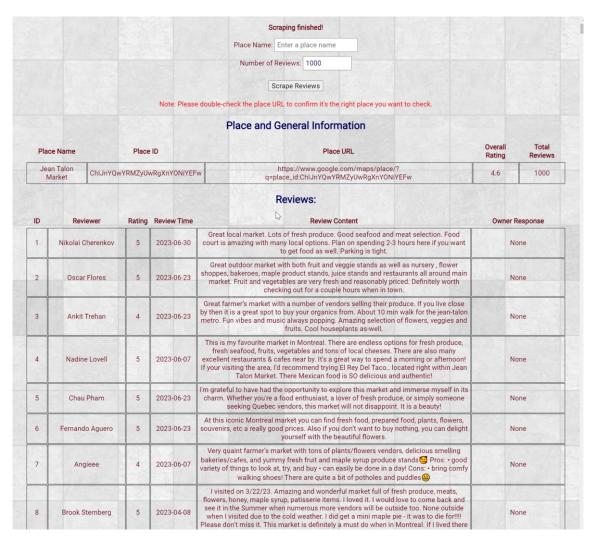
6) Input the place name which you want to scrape reviews for and the number of reviews you want. The default value of the reviews' number would be 1000. For example, now I'm inputting "Jean Talon Market" with the default value. Once you click the "Scrape Reviews" button, the scraper would begin to work and open an auto scrolling google map webpage for "Jean Talon Market" open by Firefox brower. And in the app console, you could see the working progress of the scraper. Once it's done. The scraped review would be saved as a csv file as shown below. And they will be listed on front page as well. You could double check the Place URL to confirm it's the right place.



```
970/1000 reviews scraped, in progress...
980/1000 reviews scraped, in progress...
990/1000 reviews scraped, in progress...
1000/1000 reviews scraped, in progress...
1000/1000 reviews scraped, in progress...

Reviews exported to output_data\JeanTalonMarket_4.6_1000_Reviews.csv
```





7) As you could see in app console, this place has 24068 reviews. If you want to scrape as many reviews as possible, you could input a large number, for example in this case, you could input 30000. However, google seems to have a limitation for the scroller bar to load more reviews at a certain point, for example, in the following case it stopped at 1580 reviews, so our scraper will stop trying and save 1580 reviews.

```
## app *

1498/24uo8 reviews scraped, in progress...

1500/24068 reviews scraped, in progress...

1510/24068 reviews scraped, in progress...

1520/24068 reviews scraped, in progress...

1530/24068 reviews scraped, in progress...

1540/24068 reviews scraped, in progress...

1550/24068 reviews scraped, in progress...

1550/24068 reviews scraped, in progress...

1570/24068 qeviews scraped, in progress...

1580/24068 reviews scraped, in progress...

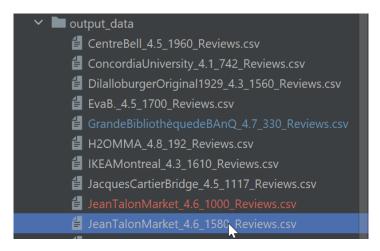
Scraping stopped due to no progress. 1580 reviews scraped.

1580/24068 reviews scraped, done.

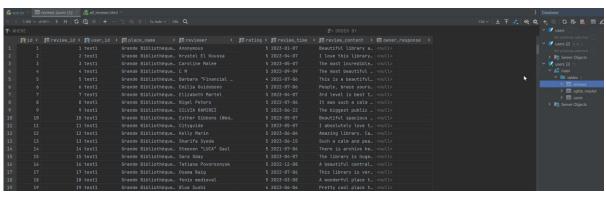
Reviews exported to output_data\JeanTalonMarket_4.6_1580_Reviews.csv

127.0.0.1 - [07/Jul/2023 00:26:27] "POST / HTTP/1.1" 200 -

127.0.0.1 - [07/Jul/2023 00:26:27] "GET /static/img/TileTex.png HTTP/1.1" 304 -
```



8) Every time you finish scraping, the scraped reviews will be saved in our database. You could check them using "All Reviews" button and "/all_reviews" page. You could delete the reviews you don't want from there. Or you could check or delete the reviews from the database directly.





3. OTHER NOTES

- 1) The scraping and scrolling are all auto but sometimes it might get stuck at a certain point, for the picture or response from owner taking up too much space in different browser/screen ratio. In this case you could try to scroll it a little bit manually to help it load more reviews. If not, the scraper will stop loading after a certain amount of trial and save the current scraped reviews.
- 2) To fit more screen size, we have set the ratio of scraper Firefox browser ratio as 50%. You could customize it for your own preference.

```
# Set the browser's zoom level to 50%

firefox_options.set_preference("layout.css.devPixelsPerPx", "0.5")
```

Please check our repo for more details: https://github.com/raeDV/Google-Reviews-Scraper