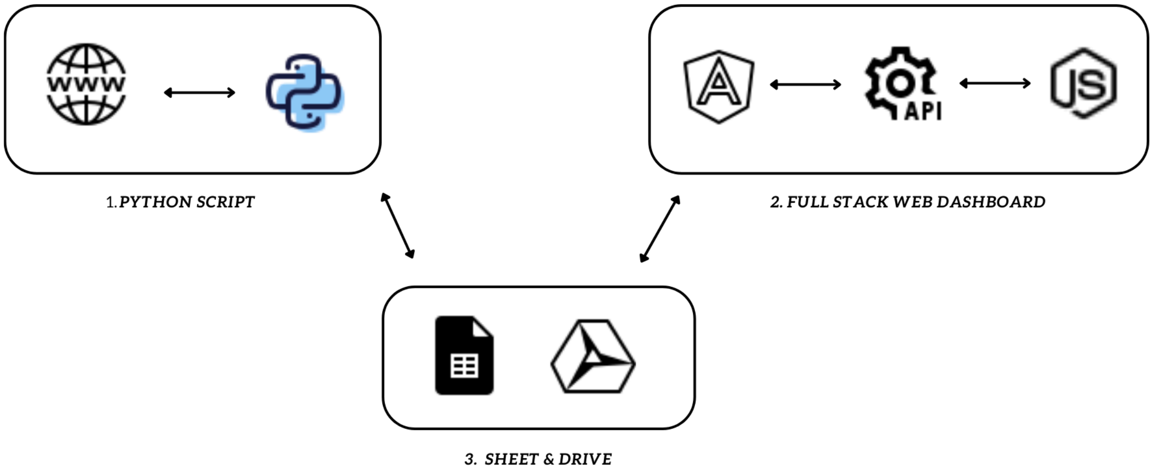
Technical Overview



1. Python Script

The web scraping script uses proxy rotation and timeouts to avoid IP blocks. Once it runs, it searches through the specified websites to find .pdf files, which it then downloads and uploads to the drive. After that, it adds detailed records to a sheet template.

2. Full Stack Web Dashboard

This is the admin panel. The frontend is built with Angular, and the backend is implemented using Node.js. It communicates through a REST API. Users can view all reports, access recently uploaded PDFs, and add new website domains to the Python script.

3. Sheet and Drive

The sheet contains detailed records about the PDFs uploaded to the drive. The drive holds the PDF files.

Installations

Unzip the delivery files. You will find two folders, 'Industry reporter' and 'Industry-reporter-app.' After that, do the following.

Python script

1. Download and install Python version 3.8.10 : visit , *Important- make sure to check the box to add Python to the system PATH.
2. Install google chrome web browser
3. Open a command prompt and type the following:
   a. pip install requests
   b. pip install beautifulsoup4
   c. pip install selenium
   d. pip install webdriver-manager
   e. pip install --upgrade google-api-python-client google-auth-httplib2 google-auth-oauthlib

Full Stack Web Dashboard

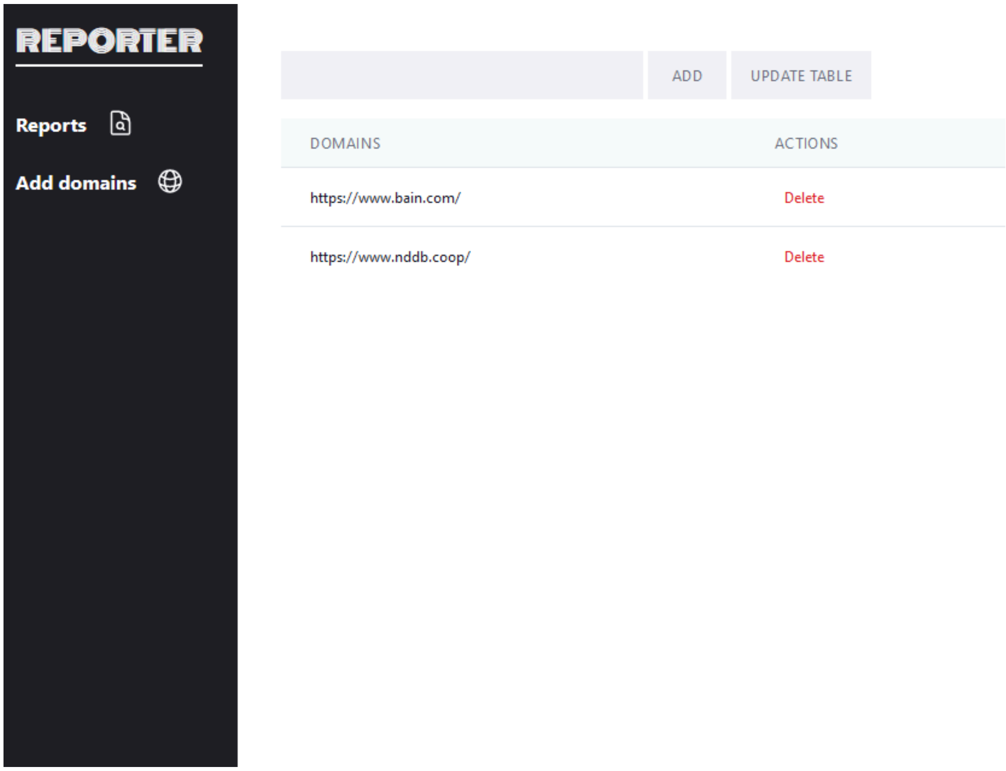
1. Download and install Node.js version 20 LTS : visit
2. Open a command prompt and type the following:
   a. npm install express
   b. npm i cors
   c. npm i google-spreadsheet --save
   d. npm i google-auth-library
   e. npm install googleapis@105 @google-cloud/local-auth@2.1.0 --save
   f. npm install -g @angular/cli
   g. Set-ExecutionPolicy -Scope CurrentUser -ExecutionPolicy RemoteSigned
3. Right-click on the 'Industry-reporter-app' folder, then choose the 'Open in Terminal' option. Type 'npm install’

Functionality and usage

Full Stack Web Dashboard

Right-click on the 'Industry-reporter-app' folder, then choose the 'Open in Terminal' option. Type 'npm run app.' This will load the Node.js server with the Angular frontend. Now go to your browser and enter: http://localhost:5000. You now have access to the admin panel.

Add domains

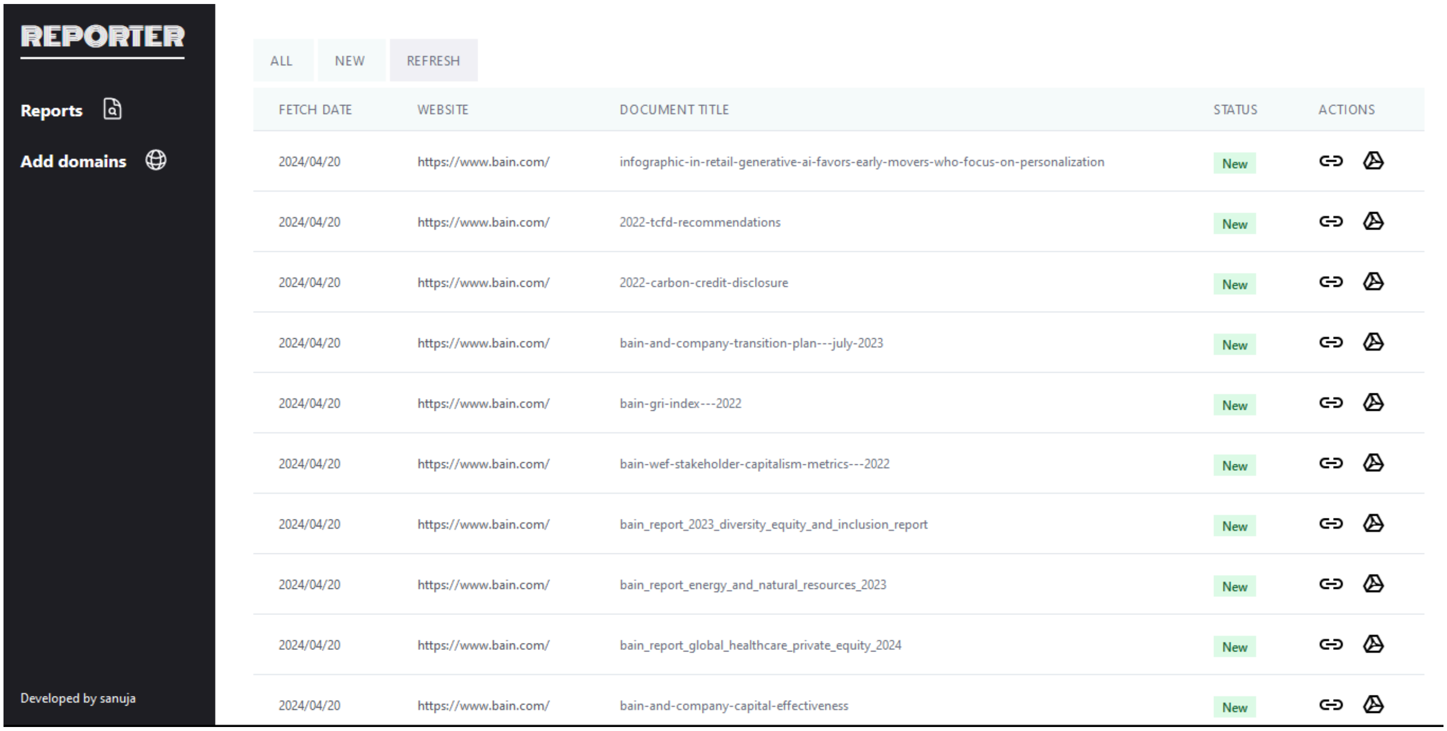


DOMAINS	ACTIONS
https://www.bain.com/	Delete
https://www.nddb.coop/	Delete

Click on 'Add Domains.' Using the 'Add' and 'Delete' buttons, create the domain sequence you want. The Python script will scrape based on this given sequence. For example, in the above example, the script will first check 'https://www.bain.com/' and then 'https://www.nddb.coop/' and so on. Once you are done with the sequence, click on the 'Update Table' button. This will save the domain sequence.

Important: If you want to add bain.com, you must enter the URL as https://www.bain.com/ and so on.

Reports



FETCH DATE	WEBSITE	DOCUMENT TITLE	STATUS	ACTIONS
2024/04/20	https://www.bain.com/	infographic-in-retail-generative-ai-favors-early-movers-who-focus-on-personalization	New	🔗 △
2024/04/20	https://www.bain.com/	2022-tcfd-recommendations	New	🔗 △
2024/04/20	https://www.bain.com/	2022-carbon-credit-disclosure	New	🔗 △
2024/04/20	https://www.bain.com/	bain-and-company-transition-plan---july-2023	New	🔗 △
2024/04/20	https://www.bain.com/	bain-gri-index---2022	New	🔗 △
2024/04/20	https://www.bain.com/	bain-wef-stakeholder-capitalism-metrics---2022	New	🔗 △
2024/04/20	https://www.bain.com/	bain_report_2023_diversity_equity_and_inclusion_report	New	🔗 △
2024/04/20	https://www.bain.com/	bain_report_energy_and_natural_resources_2023	New	🔗 △
2024/04/20	https://www.bain.com/	bain_report_global_healthcare_private_equity_2024	New	🔗 △
2024/04/20	https://www.bain.com/	bain-and-company-capital-effectiveness	New	🔗 △

Newly uploaded PDFs will have a status of 'New' in the table. Once you click the drive icon in the actions column, the 'New' status will be removed. Click the 'NEW' button at the top to filter and display only new PDFs. 'ALL' will show all the PDFs. By clicking 'REFRESH,' you can refresh the table. This 'REFRESH' function will load new data through the REST API

Python script

There are two scripts: 'first.py' and 'second.py' . Right-click on the 'Industry-reporter' folder, then choose the 'Open in Terminal' option.

Type 'python first.py.' This will find new proxies on the web and save those proxies to a JSON file. Wait until it completes. Once it does, it will display the message '[*] Run the 'second' script >>'.

If you see the message, type 'python second.py.' It will ask the following question: 'Do you want to check all the websites from the start?' The script will behave differently based on your answer.

If you type 'Y,' the script will start checking all the websites from the first URL, which is the website domain. For example, let's say a network error occurred and the script stopped working. In this case, there is no point in checking the website from the beginning; script needs to start scraping from the URL where it stopped. In cases like these, type 'N' to the question instead of 'Y'

Important: At some points, you might feel the script is a bit slow due to downloading and uploading. However, once the script has gone through every website, the issue should resolve because of the database's PDF URL tracking.