Webscraper framework

It is not permitted to use on webpages that are not allowing webscraping. This can be checked on most sites with a /robot.txt after. An example is https://www.youtube.com/robots.txt No robot has permission to get all data from that is under User-agent: * (* = meaning all).

Table of Contents

Framework Structure

This framework has the following structure from the git repo, before installation to a existing project.

- `server.ts`: Where all the API called happen.
- `routes`/: This directory contains all route definitions.
- `routes/utils/helper` Where all functions used in routes is located.
- `hooks`/: Where the premade connection to the server form making it easier for the developer is.
- `components/` - Where the demo file is.
- `components/ScraperDemo.ts` - Is a demostration on how to use this framework.
- `test-next/` - This folder contains a next.js project that is using this framework. This could be just as a test for the framework with no need of coding, just the prerequisits mention in the use the existing test project
- `server.test.ts` - Where the test enviroment is located
- `install-webscraper.js` - Is the script that is adding this framework to the developers project.
- `README.md` - Is this current file/ containing getting started/ installation etc...
- `README.pdf` - PDF version of README.md

Getting Started

This installation guide is for after the user has created a React or Next.js project.

This must be installed first before the installation guide

- Node.js - This is also needed for React or Next.js to work aswell.
- Git - Otherwise all commands may not work. This is also for making it easier to pull data from my github repository since the web scraper framework is installed in a similar way as React or Next.js framework.
- To check if both is installed open a terminal/cmd/powershell and type `git -v` and `node -v` and the result of each command should be what version you have installed

Installation of the framework

1. Run `npm i Bass4Nation/webscraper` in the project directory for downloading the framework from my github.
2. Run `npx install-webscraper` this will merge package.json with webscrapers package.json. Whis will also copy and make folders needed for the webscraper to work. This is the install-webscraper.js file. Installation similar to React or Next.js.
3. Check out How to run the the demo for the webscraper if you want examples on how to use this framework.

How to run the demo for the webscraper

The demo is just an example on how to use the framework, and can be a help on how the developer want to use it and maybe make it easier if the developer want it to do something else.

As a new project

This is how I would done it at a start next js project. (ps: This is just a example on where to place the demo)

1. In `pages/index.tsx`.
   1. `import "ScraperResult" from '@/components/ScraperResult.tsx'`
   2. Add `<ScraperDemo />` somewhere in the frontend in the file. (Must be in the front end web part of the project.)
   3. In terminal type `npm run dev`
   4. In the terminal it will say that start-server and start-client has started at port 3000 and 3002.
   5. Just copy a url from any pages. Ex: https://b4n.no/ it is okey to use since it is my portfolio page and I allow the use of a scraper on that page.

Use the existing test project in github repository.

This was used for developing this framework and is made for easy test of this framework. I will make sure it has the latest version of the framework.

1. Clone the project from github instead of using npm install.
   1. https://github.com/Bass4Nation/webscraper - Webscraper's Github repository that would be needed to for cloning
2. In a terminal navigate to `test-next` folder.
3. Run `npm install` to install the dependencies.

4. Just run `npm run dev` to start the server and the website.
5. On the website launched `localhost:3000`, paste the URL of the website you want to scrape and click "Scrape". The scraped data will be displayed on the page/console and on the server at this moment. (This is just a test to see if the scraping works. And an example on how to use the scraper.) Could be a good idea to use a page that is not protected by a robot.txt file. Ex: https://b4n.no/ it is okey to use since it is my portfolio page and I allow the use of a scraper on that page.

PS: If getting error something like this:

```
scrapescreenshotlist.ts:
Error: ENOENT: no such file or directory, scandir './public/scraped-screenshots'
    at Object.readdirSync (node:fs:1438:3)
    at C:\Users\USER\Desktop\webscraper-main\test-next\routes\utils\helper.ts:68:22
    at Generator.next (<anonymous>)
    at C:\Users\USER\Desktop\webscraper-main\test-next\routes\utils\helper.ts:8:71
    at new Promise (<anonymous>)
    at __awaiter (C:\Users\USER\Desktop\webscraper-main\test-next\routes\utils\helper.ts:4:12)
    at screenshotList (C:\Users\USER\Desktop\webscraper-main\test-next\routes\utils\helper.ts:64:61)
    at C:\Users\USER\Desktop\webscraper-main\test-next\routes\scrapescreenshotlist.ts:19:40
    at Generator.next (<anonymous>)
    at C:\Users\USER\Desktop\webscraper-main\test-next\routes\scrapescreenshotlist.ts:8:71 {
  errno: -4058,
  syscall: 'scandir',
  code: 'ENOENT',
  path: './public/scraped-screenshots'
}
```

A fix is to install the webscraper again to test-next folder. (npm i Bass4Nation/webscraper) and after npx install-webscraper again. (If the folders are missing in public folder. Don't know why this happens. Maybe it is because I don't want images to be uploaded to github so I have added it to gitignore.)

Some usage of the framework

1. Paste the URL of the website you want to scrape in the input field.
2. Click "Scrape".
3. The scraped data will be displayed on the page/console and on the server at this moment.
4. You can also download the scraped data as a JSON file. (Not yet implemented)

Additional notes

This framework may be updated after school delivery date 1.June 2023. But it will be a waiting periode because this is a school assignment/exam, so I want it to be graded first.

Where to find the framework after downloading to your project?

This is after `npm i Bass4Nation/webscraper`. All packages/frameworks will be installed in `node_modules/` folder and will be named `webscraper`

What happens when running `npx install-webscraper`?

It will copy all files that are required for the framework from `node_modules/webscraper` to root level of your project and also create all folders needed for the existing webscraper framework. Also merging this framework's package.json with the projects current package

How to run just the website or just the server?

For running just the website run `npm run start-client` and for just the server run `npm run start-server`. This is just for testing purposes and not for production.

How to run the test enviroment?

1. Run `npm test` in the root folder of the project. This will start the test enviroment. This test are just for testing the framework and not for testing the developer's code. When installing the framework this test enviroment will not be installed. This is just for testing the framework itself in the Github repository.

Yet to be implemented

1. Arrange the scraped data in a more readable format.
2. Possible to use this in other than Node.js/Next.js environments.
3. Make the framework more dynamic for the developer. Like scrapeproduct, now it is set in stone on how to to scrape, so the developer must make their store scrape function similiar to the existing but adjusting it to the store they want to scrape from.
4. Uninstalling the framework from a current project. Another way then just deleting files