Webscraper

It is not allowed 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: *.

1 Getting Started

Note: This is a work in progress. Some elements may not be implemented yet.

1.1 Installation

(ps: This is after installation of Next.js with command `npx create-next-app projectnamehere or .` for this folder.)

1. Install Node.js and install Git. Otherwise all commands may not work. You can check if you have them already installed with `git -v` and `node -v` in a terminal.
2. Run `npm install Bass4Nation/webscraper` in the project directory for downloading and installation some of the elements used
3. 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.

See Run demo of webscraper for how to use it

4. Run `npm run dev` to start the server and the website.
5. On the website, 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.)

1.2 How to run the demo for the webscraper

1.2.1 As a new project

This is how I would done it at a start next js project.

1. In pages/index.tsx.
   (a) `import "ScraperResult" from '/components/ScraperResult.tsx'`
   (b) Add `<ScraperResult />` somewhere in the frontend in the file.
   (c) In terminal npm run dev
   (d) In the terminal it will say that start-server and start-client has started at port 3000 and 3002.
   (e) 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.

1.2.2 Use the existing test project.

1. Clone the project from github instead of using npm install.
   (a) https://github.com/Bass4Nation/webscraper - Webscraper's Github repository that would be needed to for cloning
2. In a terminal navigate to tttest-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, 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.

1.3 Usage

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)

1.4 Yet to be implemented

1. Downloading the scraped data as a JSON file.
2. Arrange the scraped data in a more readable format.
3. Possible to use this in other than Node.js/Next.js environments.

Prerequisite: git, node, terminal, a code editor (I used Visual studio code to create this framework)

```
npm install Bass4Nation/webscraper
npx install-webscraper
npm run dev
```

Server and client on separate port. Usually server on 3002 and client is 3000.

For testing the demo that is in the code. You must import ScraperResult.tsx Can be shown in files at webscraper github how to import it and use it

```
import ScraperResult from '@/components/ScraperResult.tsx'
```