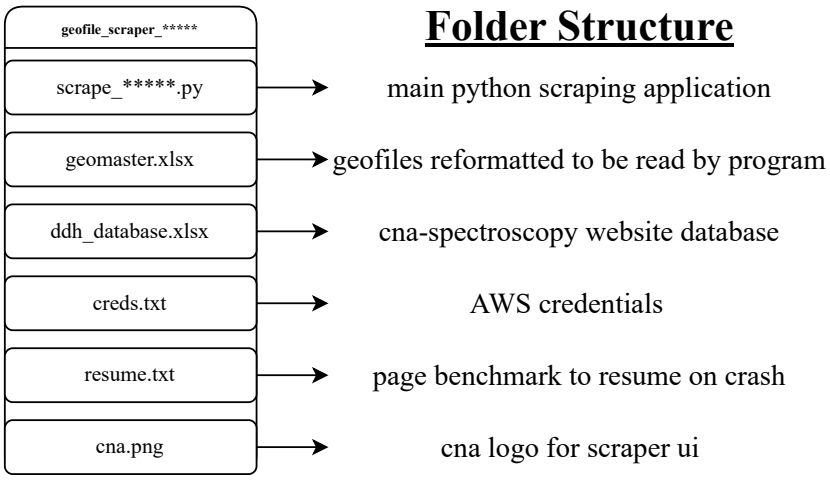
Geofile Webscraper

How to Guide



Folder Structure

Before starting

1. Download any python terminal.

2. In your terminal install the required packages:

```
pip install requests          #http requests
pip install beautifulsoup4    #scrapes html and xml files
pip install lxml              #xml file processing
pip install boto3             #direct access to S3
pip install openpyxl          #read/write excel files
```

3. Playwright install, you can use either option:
    1. pip install playwright #python
    2. npm install playwright #node
    3. VSCode extension

Starting the scraper

1. Open creds.txt, input your AWS login credentials in this format "accesskey, secretkey"

2. Open resume.txt and make sure its empty.

3. Using any python terminal change to the working directory of the scraper. ex:
"cd /d C:\Users\marcus.bourne\Desktop\scraping tool\geofiles_scraper_s3"

4. Run the script using "python scrape_*****.py"

5. When the UI opens, select "Start Scraper" to begin.

6. If the scraper runs into any issues, it will automatically restart on itself

Notes

1. If you wish to begin on a specific page, open resume.txt and input where you would like to jump to.

2. To upload files to s3 run scrape_s3.py

3. To save files locally run scrape_local.py

4. To scrape and download everything on the Mines and Energy website run scrape_all.py

5. Current S3 bucket path "cna-webfiles/Textract/input". To save somewhere different edit lines 27/28 in the code.