亞洲大學
ASIA UNIVERSITY

Midterm Project Report

Advanced Computer Programming

Student Name : Chinzorig Battulga

Student ID : 113021191

Teacher : DINH-TRUNG VU

2025-04

Chapter 1 Introduction

1.1 Github

1) **Personal Github Account**: https://github.com/Mimikooo
2) **Group Project Repository**: https://github.com/Jantsagdorj/ACP-AU-1132

The objective of this project was to create a web scraper using Python's Scrapy framework to extract structured information from a GitHub profile. This tool is particularly useful for developers and researchers who wish to analyze repositories at scale without manually copying data. The main goal was to collect information from each public repository under a specific user, including its URL, description (About), last updated timestamp, programming languages used, and number of commits.

1.2 Overview

To complete this project, the following advanced Python programming tools and libraries were used:

- **Scrapy**: The core framework used for web scraping. It allowed for easy spider creation, request handling, and HTML data extraction.
- **cssselect**: Used in conjunction with Scrapy to define CSS selectors and extract specific elements from the GitHub HTML structure.
- **Feed exporter**: A Scrapy built-in tool that outputs scraped data in structured formats, such as XML, JSON, and CSV. For this project, XML was chosen.

The scraper navigates from the GitHub profile page (https://github.com/Mimikooo?tab=repositories) to each listed repository, gathering relevant data. The collected data is exported to an XML file called repos.xml, which contains detailed information for further use.

Implementation

1.1 Setup and Environment

1.1.1 Environment Information

- Operating System: Windows 11
- Python Version: 3.12
- Scrapy Version: 2.12.0
- IDE/Text Editor: VS Code and PowerShell Terminal

1.1.2 Installing ScrapyFunctions and creating the project

To install Scrapy, I used pip in the PowerShell terminal:

```
PS C:\Users\chinz> pip install scrapy
Defaulting to user installation because no
```

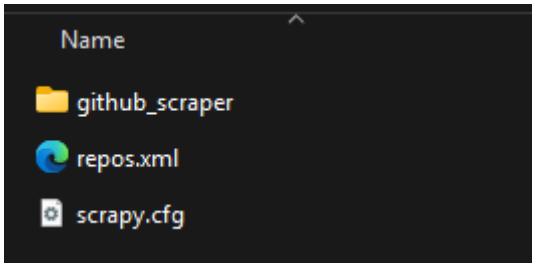
After installation, I verified that Scrapy was working:

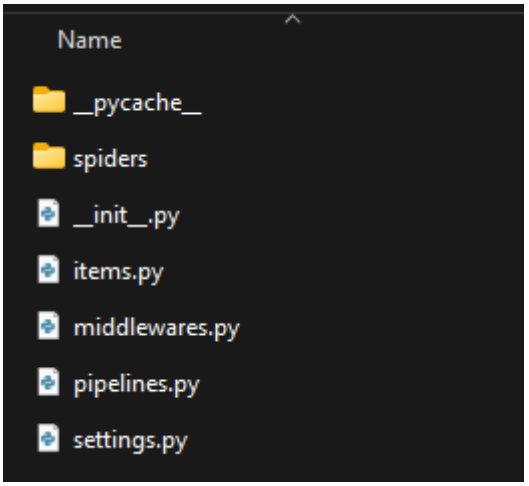
```
PS C:\Users\chinz\OneDrive\Documents\Advanced Computer - Midterm\github_scraper> python -m scrapy -v
Scrapy 2.12.0 - active project: github_scraper
```

I created a new Scrapy project named github_scraper with the following command:

```
PS C:\Users\chinz\OneDrive\Documents\Advanced Computer - Midterm> python -m scrapy startproject github_scraper
New Scrapy project 'github_scraper', using template directory 'C:\Users\chinz\AppData\Local\Packages\PythonSoftwareFound
ation.Python.3.12_qbz5n2kfra8p0\LocalCache\local-packages\Python312\site-packages\scrapy\templates\project', created in:
    C:\Users\chinz\OneDrive\Documents\Advanced Computer - Midterm\github_scraper
```

This generated the necessary folder structure:





1.2 Class

Class: GithubScraperItem (items.py)

Description: Defines the fields that will be scraped and stored for each repository. This item acts like a data structure.

Fields:

- url: Complete URL to the GitHub repository
- about: Description or purpose of the repository. If not provided and the repository is not empty, defaults to the repository name.
- last_updated: Datetime string showing the latest update timestamp
- languages: List of programming languages used in the repository (if any)
- commits: Total number of commits in the repository (if available)

items.py 1

C: > Users > chinz > OneDrive > Documents > Advanced Computer - Midterm >

```python
import scrapy

class GithubScraperItem(scrapy.Item):
    url = scrapy.Field()
    about = scrapy.Field()
    last_updated = scrapy.Field()
    languages = scrapy.Field()
    commits = scrapy.Field()

```

1.3 Spider

Spider: github_spider.py

Description: Scrapy spider that starts from the repositories page and navigates into each repo to extract information.

Method: parse()

Navigates through the repositories listed on the user's GitHub page and follows each repository link.

Method: parse_repo(response)

Scrapes details of each repository:

- Checks if the repo is empty
- Extracts about, last_updated, languages, and commits (if available)
- Yields structured GithubScraperItem

Method: Get repo details from API

```
repo_api_url = f"https://api.github.com/repos/{owner}/{repo_name}"
repo_data = requests.get(repo_api_url, headers=headers).json()
```

- Purpose: This calls the GitHub REST API to get detailed info about a specific repository.
- repo_api_url is constructed using the owner and repo_name.
- The requests.get(...) sends a GET request to the GitHub API.
- .json() converts the API response to a Python dictionary (repo_data).

Method: Get last_updated

```
item['last_updated'] = repo_data.get('updated_at')
```

- Purpose: Stores the repository's last update timestamp into your item (Scrapy item).
- This uses the updated_at field from the API response.

Method: Get programming languages

```
lang_url = f"https://api.github.com/repos/{owner}/{repo_name}/languages"
lang_data = requests.get(lang_url, headers=headers).json()
item['languages'] = list(lang_data.keys()) if lang_data else None
```

- Purpose: Fetches a breakdown of programming languages used in the repo.
- Each language is a key in the lang_data dictionary. list(lang_data.keys()) returns just the language names.

C: > Users > chinz > OneDrive > Documents > Advanced Computer - Midterm > github_scraper > github_scraper > spiders > github_spider.py > GithubSpider > parse_repo

```python
import scrapy
import requests
from github_scraper.items import GithubScraperItem

class GithubSpider(scrapy.Spider):
    name = "github"
    allowed_domains = ["github.com"]
    start_urls = ["https://github.com/Mimikooo?tab=repositories"]

    Tabnine | Edit | Test | Explain | Document
    def start_requests(self):
        for url in self.start_urls:
            yield scrapy.Request(url=url, callback=self.parse, headers={
                'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36'
            })

    Tabnine | Edit | Test | Explain | Document
    def parse(self, response):
        repo_links = response.css('h3 a::attr(href)').getall()
        for link in repo_links:
            repo_url = response.urljoin(link)
            yield scrapy.Request(url=repo_url, callback=self.parse_repo, headers={
                'User-Agent': 'Mozilla/5.0'
            })

    Tabnine | Edit | Test | Explain | Document
    def parse_repo(self, response):
        item = GithubScraperItem()
        item['url'] = response.url

        about = response.css('p.f4::text, p.f4 span::text').get()
        about = about.strip() if about else None
        is_empty = response.css('.blankslate').get() is not None

        repo_name = response.url.split('/')[-1]
        owner = response.url.split('/')[-2]
        item['about'] = about if about else (None if is_empty else repo_name)

        headers = {
            'Accept': 'application/vnd.github.v3+json',
            'User-Agent': 'Mozilla/5.0',
        }

        # 1. Get repo details from API
        repo_api_url = f"https://api.github.com/repos/{owner}/{repo_name}"
        repo_data = requests.get(repo_api_url, headers=headers).json()

        # 2. last_updated
        item['last_updated'] = repo_data.get('updated_at')
```


```python
        # 1. Get repo details from API
        repo_api_url = f"https://api.github.com/repos/{owner}/{repo_name}"
        repo_data = requests.get(repo_api_url, headers=headers).json()

        # 2. last_updated
        item['last_updated'] = repo_data.get('updated_at')
```


```python
        # 3. languages
        lang_url = f"https://api.github.com/repos/{owner}/{repo_name}/languages"
        lang_data = requests.get(lang_url, headers=headers).json()
        item['languages'] = list(lang_data.keys()) if lang_data else None

        # 4. commits
        commits_url = f"https://api.github.com/repos/{owner}/{repo_name}/commits?per_page=1"
        response_api = requests.get(commits_url, headers=headers)

        if 'Link' in response_api.headers:
            links = response_api.headers['Link']
            for link in links.split(','):
                if 'rel="last"' in link:
                    last_url = link[link.find('<')+1:link.find('>')]
                    page_number = int(last_url.split('page=')[-1])
                    item['commits'] = page_number
                    break
            else:
                item['commits'] = 1
        else:
            item['commits'] = len(response_api.json())

        yield item
```

Chapter 2 Results

1.1 Results

Result 1: The spider successfully scraped the public repository Test. Since this repository is likely empty or minimal, some fields returned None.

Output (repos.xml):

```
<item>
  <url>https://github.com/Mimikooo/Test</url>
  <about>Test</about>
  <last_updated>None</last_updated>
  <languages>None</languages>
  <commits>None</commits>
</item>
```

Result 2: This time the repository is not empty and gets all the data using the scrapy web crawler.

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<items>
  ▼<item>
    <url>https://github.com/Mimikooo/Test</url>
    <about>This is the about section I just added at 10:49AM 4/14/2025</about>
    <last_updated>2025-04-14T03:22:34Z</last_updated>
   ▼<languages>
      <value>Python</value>
    </languages>
    <commits>8</commits>
  </item>
</items>
```

Chapter 2 Conclusions

This project successfully demonstrated how to combine Scrapy with the GitHub API for accurate and structured data scraping. By using API endpoints instead of relying solely on HTML, the scraper was able to retrieve reliable information like update timestamps, languages, and commit counts.

The updated approach improved accuracy, avoided issues with dynamic content, and laid a strong foundation for building more advanced data collection tools.