

****

**Midterm Project Report**

**Advanced Computer Programming**

**Student Name : Bat-Erdene Tsogbat**

**Student ID : 113021197**

**Teacher : DINH-TRUNG VU**

**2024-04**

# Introduction

## Github

1. **Personal Github Account**: https://github.com/Tsogbat
2. **Group Project Repository**: https://github.com/113021197/ACP-Group11

## Overview

For this project I used “Scrapy” to scrape GitHub user profile and it’s repositories. “Scrapy” is a powerful Python framework that is used for web data extraction, web crawling. For this report I will use the term “web data extraction” and “web scraping” interchangibly. This scraper uses GitHub’s REST API to fetch accurate data’s from all public repositories of a given user.

This projects main goal is to scrape exact informations of:

* Repository description (about)
* Last updated time (last\_updated)
* Programming languages and their percentages in repository (languages)
* And lastly the total number of commits (number\_of\_commits)

The main Advanced libraries used for these are:

* Scrapy: Framework for web data extraction
* Json: built-in python library used for parsing API responses

# Implementation

## Class

## GithubRepoSpider

This class is the main class that contains and defines the Scrapy’s spider logic. It handles crawling, API request and also the json responses.

### Fields

* Name: giving the spider name “github\_repos”
* Start\_urls: url that we need to extract information from.
* Custom\_settings: for customizing the settings. When scrapy is used on local machine, or code editor or IDE we can use the cmd command “scrapy startproject (project\_name)” and that will give the initial configurations. So this field allows us to mimic that and also gives us the freedom of custom configurations such as XML output format and user agent.

### Methods

* Parse: it parses the GitHub profile page and extracts the all public repository links.
* Parse\_api\_repo: handles the main repository metadata from GitHub’s API, and checks if the repo is empty.
* Parse\_languages: it collects the programming language used in repository and calculates the percentage.
* Parse\_commits: it collects the total number of commits.

### Functions

* Json.load for json data extraction
* Urljoin for url building

## Class 2

## CrawlerProcess

This class is used to run the scrapy spider and basically runs the whole web scraping code.

# Results

## Result

.Here is the result of web scraping.

<?xml version="1.0" encoding="utf-8"?>

<items>

<item><url>https://github.com/Tsogbat/tachiyomi</url><about>Manga/Manhwa reading app</about><last\_updated>2024-01-16T06:27:00Z</last\_updated><languages><value>Swift (100.0%)</value></languages><number\_of\_commits>20</number\_of\_commits></item>

<item><url>https://github.com/Tsogbat/PersonalWeb</url><about>PersonalWeb</about><last\_updated>2025-03-11T01:51:36Z</last\_updated><languages><value>TypeScript (95.72%)</value><value>CSS (3.74%)</value><value>JavaScript (0.34%)</value><value>MDX (0.2%)</value></languages><number\_of\_commits>5</number\_of\_commits></item>

<item><url>https://github.com/Tsogbat/responsive-navigation-menubar</url><about>Responsive Navigation Menu Bar Using HTML CSS &amp; JavaScript</about><last\_updated>2024-01-13T03:22:36Z</last\_updated><languages><value>CSS (38.23%)</value><value>SCSS (36.97%)</value><value>HTML (19.0%)</value><value>JavaScript (5.8%)</value></languages><number\_of\_commits>7</number\_of\_commits></item>

</items>

## Enhanced result

1st repository

* URL: https://github.com/Tsogbat/tachiyomi
* About: Manga/Manhwa reading app
* Last updated: 2024-01-16T06:27:00Z
* Languages: Swift (100.0%)
* Commits: 20

2nd repository

* URL: https://github.com/Tsogbat/PersonalWeb
* About: PersonalWeb
* Last updated: 2025-03-11T01:51:36Z
* Languages: TypeScript (95.72%) CSS (3.74%) JavaScript (0.34%) MDX (0.2%)
* Commits: 5

3rd repository

* URL: https://github.com/Tsogbat/responsive-navigation-menubar
* About: Responsive Navigation Menu Bar Using HTML CSS & JavaScript
* Last updated: 2024-01-13T03:22:36Z
* Languages: CSS (38.23%) SCSS (36.97%) HTML (19.0%) JavaScript (5.8%)
* Commits: 7

# Conclusions

So in this project, I developed GitHub user profile scraper using the Scrapy framework in python. The goal was to extract exact informations from a given user’s profile such as repository URL, description, last updated date, what programming languages where used, and the total number of commits. This program also uses the GitHub’s API to fetch accurate datas from repositories.

The final output was given in XML and included key datas that usually stored in repositories. As you can see, each repository has different main programming languages and every repository was checked by hand ensuring the accuracy.

This program highlights the use of advanced computer programming techniques and showed me how basic web scraping works. This simple program gave me foundations of big data and data processing, data analysis, advanced computer programming and future of computer science.