

****

**Midterm Project Report**

**Advanced Computer Programming**

**Student Name : Giandra Raihan Nayottama**

**Student ID : 113021219**

**Teacher : DINH-TRUNG VU**

**2025-04**

# Introduction

## Github

1. **Personal Github Account**: https://github.com/nayottama11
2. **Group Project Repository**: https://github.com/wonderhorse90/ACP

## Overview

The program uses data classes to structure repository data, pprint for readable XML output, and pattern matching to streamline data processing. Regular expressions are used to extract specific information from the page content, while Scrapy handles the web scraping and navigates multiple pages. As a result, the program extracts the main details of repositories, including URL, About, Languages, and Commit count, and stores the data in an XML file.

# Implementation

## Class 1

GitHubSpider

### Fields

* **name:** Spider's unique identifier.
* **allowed\_domains:** Domains allowed to crawl.
* **start\_urls:** Initial URL to begin crawling.

### Methods

* **start\_requests：**Sends the first request to start\_urls.
* **parse(response):** Handles parsing of the repository listing page.
* **parse\_repository\_details(response):** Extracts details from each repository's individual page.

### Functions

* **start\_requests：**Sends the first request to start\_urls.
* **parse(response):** Handles parsing of the repository listing page.
* **parse\_repository\_details(response):** Extracts details from each repository's individual page.

## Class 2

GitHubRepoItem

### Fields

* url: Repository URL.
* name: Name of the repository.
* about: Description of the repository.
* last\_updated: Date when the repository was last updated.
* languages: Programming languages used in the repo.
* commits: Number of commits in the repository.
* is\_empty: Boolean value to indicate if the repository is empty.

# Results

## Result 1

<?xml version="1.0" encoding="utf8"?>

<items>

<item><url>https://github.com/nayottama11/kanji</url><name>kanji</name><about>kanji</about><last\_updated>2024-04-25</last\_updated><is\_empty>False</is\_empty><languages><value>HTML</value><value>66.3%</value><value>JavaScript</value><value>22.6%</value><value>Jupyter Notebook</value><value>9.6%</value><value>CSS</value><value>1.5%</value></languages><commits>1</commits></item>

<item><url>https://github.com/nayottama11/Mini-Project-React</url><name>Mini-Project-React</name><about>Mini-Project-React</about><last\_updated>2023-09-12</last\_updated><is\_empty>True</is\_empty><languages>None</languages><commits>None</commits></item>

<item><url>https://github.com/nayottama11/my-profile</url><name>my-profile</name><about>test</about><last\_updated>2023-08-09</last\_updated><is\_empty>False</is\_empty><languages><value>HTML</value><value>100.0%</value></languages><commits>1</commits></item>

<item><url>https://github.com/nayottama11/ets-webprog</url><name>ets-webprog</name><about>ets-webprog</about><last\_updated>2023-10-26</last\_updated><is\_empty>False</is\_empty><languages><value>PHP</value><value>50.0%</value><value>Blade</value><value>36.3%</value><value>CSS</value><value>13.5%</value><value>JavaScript</value><value>0.2%</value></languages><commits>1</commits></item>

<item><url>https://github.com/nayottama11/quiz1</url><name>quiz1</name><about>quiz1</about><last\_updated>2023-09-28</last\_updated><is\_empty>False</is\_empty><languages><value>HTML</value><value>88.2%</value><value>CSS</value><value>11.8%</value></languages><commits>1</commits></item>

<item><url>https://github.com/nayottama11/mini-project-2</url><name>mini-project-2</name><about>mini-project-2</about><last\_updated>2023-09-12</last\_updated><is\_empty>False</is\_empty><languages><value>None</value></languages><commits>1</commits></item>

<item><url>https://github.com/nayottama11/final-project</url><name>final-project</name><about>final-project</about><last\_updated>2023-11-09</last\_updated><is\_empty>False</is\_empty><languages><value>JavaScript</value><value>56.4%</value><value>CSS</value><value>42.8%</value><value>HTML</value><value>0.8%</value></languages><commits>2</commits></item>

<item><url>https://github.com/nayottama11/react-redux</url><name>react-redux</name><about>react-redux</about><last\_updated>2023-09-03</last\_updated><is\_empty>False</is\_empty><languages><value>None</value></languages><commits>2</commits></item>

<item><url>https://github.com/nayottama11/dibimbing-students-profile</url><name>dibimbing-students-profile</name><about>dibimbing-students-profile</about><last\_updated>2023-08-09</last\_updated><is\_empty>False</is\_empty><languages><value>None</value></languages><commits>11</commits></item>

<item><url>https://github.com/nayottama11/Mini-Project-Unfinished</url><name>Mini-Project-Unfinished</name><about>Mini-Project-Unfinished</about><last\_updated>2023-08-04</last\_updated><is\_empty>False</is\_empty><languages><value>HTML</value><value>65.8%</value><value>CSS</value><value>34.2%</value></languages><commits>1</commits></item>

<item><url>https://github.com/nayottama11/quiz2</url><name>quiz2</name><about>Laravel Project</about><last\_updated>2023-11-23</last\_updated><is\_empty>False</is\_empty><languages><value>None</value></languages><commits>2</commits></item>

<item><url>https://github.com/nayottama11/final-project-webpro</url><name>final-project-webpro</name><about>final-project-webpro</about><last\_updated>2023-12-06</last\_updated><is\_empty>False</is\_empty><languages><value>HTML</value><value>61.3%</value><value>CSS</value><value>38.7%</value></languages><commits>1</commits></item>

</items>

# Conclusions

This Scrapy program helps automate the process of collecting data from a GitHub user's repositories. It grabs useful info like the repo name, description, last update, languages used, number of commits, and whether it's empty or not. The data is organized nicely using a custom item class and saved in XML format.