

****

**Midterm Project Report**

**Advanced Computer Programming**

**Student Name : Muhammad Luthfi Zafir**

**Student ID : 113021203**

**Teacher : DINH-TRUNG VU**

**2025-04**

# Introduction

## Github

1. **Personal Github Account**: https://github.com/luthfizafir
2. **Group Project Repository**:

## Overview

In this project, I made a Scrapy spider to extract structured information from a GitHub profile page. The spider collects metadata from all repositories such as URL, About section, Languages used, Last Updated date, and Number of Commits.

Advanced Python libraries and features used:

* **Scrapy** : for web crawling and parsing HTML content.
* **CSS Selectors** : to target specific elements on GitHub pages.
* **Built-in functions and dictionary operations** : for data structuring.

This script automatically navigates through each repository and adapts if the repo is empty or missing an About section. The data is exported to an XML file.

# Implementation

## GithubSpider Class

This class is responsible for crawling at the Github repositories. It also holds the name that will be used foor the XML output.

### Fields

* **Name:** identifies spider name.

### Methods

* start\_requests: Starts the crawl from the GitHub repositories tab.
* parse: Extracts the URL of each repository and schedules a new request.
* parse\_repo: Parses each individual repository's page to extract data fields needed (URL, about, programming languages, commits, last updated).

### parse\_repo Functions

* Checks if the About section exists. If not and the repository is not empty, uses the repo name as a substitute.
* Extracts the programming languages used.
* Extracts the last updated date and number of commits (if not an empty repo).

## Extract About

Collects the about data in the repository page, if there is none, it will printout the name of the project

## Extract Languages & Last Updated

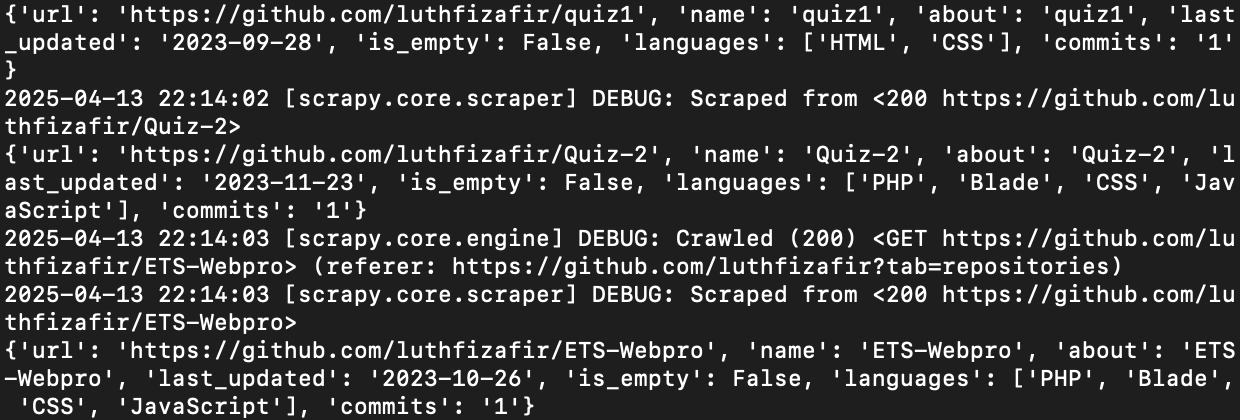
This function is to extract the languages used and the last time the project got updated

## Extract Commit

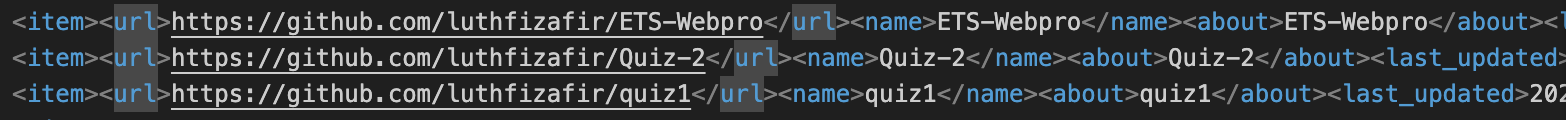
Extract the number of time the project is commited.

# Results

## Github Scrapy Result

. 

## XML Result



# Conclusions

This project demonstrates effective usage of Scrapy for web scraping and data extraction. The spider can be adapted for other GitHub profiles or even extended to include topics, stars, and forks. It also highlights the importance of error handling and fallback logic when parsing inconsistent web content.