Final Report - CodeScraper Mini Project

In this team project, we decided to go beyond a simple script and create a complete, modular system for crawling and processing classified ads. Our team had 4 members, each responsible for a specific part:

- Amin Sharifi developed the Detector module to identify new ads.
- Mahdieh Karamati implemented the Scraper module to extract details from ad pages.
- AmirAla Bonaghar handled the Cleaner module to standardize and clean data.
- AmirAla Bonaghar and Erfan Saraei designed and implemented the database.
- Erfan Saraei also developed the ad similarity detection algorithm.

Challenges:

Our biggest issue was handling site blocks from Divar. We had to be careful with the request rate and use proper timing (e.g., sleep delays) to avoid being blocked.

Architecture:

The project was structured modularly, using object-oriented design to keep everything clean, maintainable, and extendable. This helped each person focus on their part without confusion.

Improvements:

With more time, we could have improved the Scraper UI and the final display in the console to be more visually appealing.

Technologies Used:

- requests: for fetching web pages

- sqlite: lightweight database storage
- rich: for improved console output

Expandability:

Thanks to modularity and OOP principles, the system can easily support additional websites or ad categories like Services, Electronics, etc.

Real-world Usage:

To use this system in production, it only needs to be deployed on a server. Its current structure is ready for real-world application.

The project and this report are available on GitHub, and version v1.0 has been tagged in the Releases section.

We started this journey with challenges, and finished it with valuable experience.