DataRaptor
Intelligent Web Scraping and Analysis Tool

- DataRaptor is an intelligent web scraping tool that automates interactions with websites that lack APIs.
- It scrapes data, classifies it, predicts trends such as price variations, and stores the extracted information into a PostgreSQL database.

Team:Bifa111
Artur Rembas

Project Objectives

- Automate login and scraping from websites that don’t provide APIs.
- Classify scraped data into categories such as product types or data types.
- Predict price variations using machine learning models.
- Store data in PostgreSQL for further analysis and reporting.

Team:Bifa111
Artur Rembas

System Architecture

- Visual representation of the system's architecture.
- Flow: User Input → Web Scraper (Selenium) → AI (Classification, Prediction) → PostgreSQL → Result Display.
- Key technologies used: Selenium, BeautifulSoup, PostgreSQL, TensorFlow, scikit-learn.

Team:Bifa111
Artur Rembas

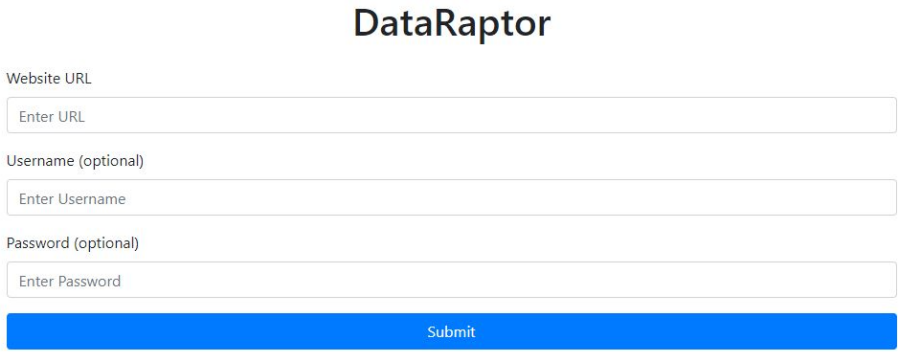
Key Features

- Login Automation: Automatically log into websites using Selenium.
- Dynamic Scraping: Adjusts scraping logic dynamically using machine learning.
- Data Classification: Categorizes scraped data using Random Forest models.
- Price Prediction: Uses a neural network to predict price trends.
- Database Integration: Stores scraped data in PostgreSQL for future analysis.

Team:Bifa111
Artur Rembas

Scraping Process Flow

- Step 1: User inputs website URL and credentials (if needed).
- Step 2: Selenium opens the browser and navigates to the page.
- Step 3: BeautifulSoup extracts data from the page’s HTML structure.
- Step 4: Data is sent for classification and stored in PostgreSQL.



Team:Bifa111
Artur Rembas

Dynamic Page Structure Adaptation

- Uses machine learning to detect changes in a website's structure.
- Automatically adjusts the scraping logic when the HTML structure changes.
- Ensures consistent data extraction even when websites update their design.

Team:Bifa111
Artur Rembas

Data Classification using AI

- Classifies scraped data (e.g., products) into categories such as Category A and Category B.
- Utilizes Random Forest models for categorization based on attributes like product name and description length.
- Visual representation: A bar chart showing the number of products per category.

Team:Bifa111
Artur Rembas

Price Prediction using Neural Networks

- Uses TensorFlow to build a neural network that predicts price variations based on historical data.
- Outputs price trends and forecasts for future changes in product prices.
- Visual representation: A line graph showing predicted vs actual price trends over time.

Team:Bifa111
Artur Rembas

PostgreSQL Integration

- Scraped data is stored in a PostgreSQL database for further analysis.
- The database structure includes tables like products with fields for name, price, description, and category.
- Data can be queried, analyzed, and exported in various formats for reporting.

Team:Bifa111
Artur Rembas

Conclusion and Next Steps

- DataRaptor is a complete solution for web scraping, data classification, price prediction, and data storage.
- Next steps:
  - Enhance CAPTCHA handling capabilities.
  - Add support for scraping more websites.
  - Integrate additional analysis and monitoring tools.
  - Build an API to allow external services to interact with the DataRaptor database.

Team:Bifa111
Artur Rembas