ID 95

Enhancing Web Scraping with Artificial Intelligence: A Review

M Weerasinghe[1#], MWP Maduranga[1], and MVT Kawya[1]

[1]Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

[#]38-bis-0024@kdu.ac.lk

Abstract

Web scraping, the process of extracting data from websites, plays a crucial role in data collection for research, analysis, and automation. However, traditional web scraping techniques face challenges such as handling dynamic websites, anti-scraping measures, and extracting structured data from unstructured web pages. In recent years, artificial intelligence (AI) has emerged as a powerful tool to enhance web scraping, offering solutions to overcome these challenges and improve data extraction efficiency and effectiveness. This review explores the application of AI techniques in web scraping, including natural language processing for information extraction, machine learning for web page classification and computer vision for web page parsing. The benefits of AI-enhanced web scraping include improved accuracy, enhanced efficiency, handling dynamic websites, and scalability. Further, there are multiple challenges with the use of AI in web scraping. Ensuring the ethical and responsible use of AI in scraping is crucial to respect privacy rights, intellectual property, and terms of service of websites. However, the ethical considerations and the need to adapt to evolving anti-scraping measures pose challenges. This review highlights the potential of AI in web scraping and emphasizes the importance of responsible and ethical practices.

Keywords: *Web scraping, Artificial intelligence, Machine learning, Natural language processing, Data extraction*