IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

Hybrid Approach for Scraping HTML within JSON Structure

Shivani D. Chaudhari[1] and Sudeshna Roy[2]
Students, Master of Computer Application[1,2]
Bharati Vidyapeeth's Institute of Management and Information Technology, Navi Mumbai, India

Abstract: *Web scraping is a technique for extracting unstructured information from a website and storing it in a structured format. Scraping is an important approach for autonomously retrieving third-party data. For example, If you want to see all of the available transportation on a specific day, you may have to scrape travel websites to gather the information you need. In this research paper, we are attempting to scrape a website that has a JSON structure but also includes an HTML format that must be scraped along with the JSON data. Doing scraping with RegEx for HTML and JSON extraction is a technique for collecting data from these types of websites. To scrape JSON data, we must first get a JSON response and scrape data using key value pairs and then use RegEx to retrieve information from HTML contained within that JSON structure.*

Keywords: JSON, RegEx, Web scraping, Hybrid scraping, scrapy, HTML scraping, Data extraction

REFERENCES

[1] Zhao, Bo. "Web scraping." Encyclopedia of big data (2017): 1-3.
[2] Gunawan, R., Rahmatulloh, A., Darmawan, I., & Firdaus, F. (2019, March). Comparison of web scraping techniques: regular expression, HTML DOM and Xpath. In International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018) Comparison (Vol. 2, pp. 283-287).
[3] Chapagain, A. (2019). Hands-On Web Scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, RegEx, and others. Packt Publishing Ltd.
[4] Morsidi, F., Sulaiman, S., Wahid, R. A., & Malim, T. (2017). Feature extraction using regular expression in detecting proper noun for Malay news articles based on KNN algorithm. Journal of Fundamental and Applied Sciences, 9(5S), 210-231.
[5] Uzun, E. (2020). A novel web scraping approach using the additional information obtained from web pages. IEEE Access, 8, 61726-61740.
[6] Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. Knowledge-based systems, 70, 301-323.
[7] Uzun, E., Agun, H. V., & Yerlikaya, T. (2013). A hybrid approach for extracting informative content from web pages. Information Processing & Management, 49(4), 928-944.
[8] Uzun, E., Serdar Gu¨ner, E., Kılıc¸aslan, Y., Yerlikaya, T., & Agun, H. V. (2014). An effective and efficient Web content extractor for optimizing the crawling process. Software: Practice and Experience, 44(10), 1181- 1199.
[9] Wu, Y. C. (2016). Language independent web news extraction system based on text detection framework. Infomation Sciences, 342, 132-149.
[10] Kumar, M., Bhatia, R., & Rattan, D. (2017). A survey of Web crawlers for information retrieval. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1218.
[11] Song, D., Sun, F., & Liao, L. (2015). A hybrid approach for content extraction with text density and visual importance of DOM nodes.
Knowledge and Information Systems, 42(1), 75-96
[12] Qureshi, P. A. R., & Memon, N. (2012). Hybrid model of content extraction. Journal of Computer and System Sciences, 78(4), 1248-1257.