

Usage of Web Scrapping in the Pharmaceutical Sector

Ruby Dahiya¹, Nidhi¹, Kajal Kumari¹, Shruti Kumari¹, Nidhi Agarwal¹. *

¹ School of CSE, Galgotias University, Greater Noida, India

Abstract

INTRODUCTION: Web scraping is a technique that provides organizations with the ability to analyse large amounts of information and gather new information.

OBJECTIVES: Find a group that is a health check, a full body test, a blood test, and so on. In this way, the pharmaceutical industry should consider how to improve information, information storage, information retrieval, and capture. For example, the healthcare system may decide to standardize the assessment of speech and allow information to be shared across organizations to improve treatment outcomes in web scraping applications.

METHODS: Web scraping is based on the pharmaceutical industry. From here, we get information about pharmacies, such as drug names in different categories or drug sales. However, we are dealing with diseases and common medicines. Using this information, we can find the most common viruses. There are many factors to consider when creating a junk website for the pharmaceutical industry, such as drug names, tablet categories, and syrups found in the pharmaceutical industry.

RESULTS: As is clearly visible from the output, there are columns for drug names, manufacturers, drug types, and prices. This is the information we get from a website called Net meds, a pharmacy site. With the help of this information, we learn which drugs are most needed, and then we can find the most common diseases today.

CONCLUSION: The results of this web scraping can be very useful and powerful. However, the industry's success in web scraping and data extraction techniques depends on the availability of clean chemical data.

Keywords: Web scraping, beautiful soup, drug, medicine

Received on 25 July 2023, accepted on 30 October 2023, published on 06 November 2023

Copyright © 2023 R. Dahiya *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetpht.9.4312

*Corresponding author. Email: nidhi.agarwal@galgotiasuniversity.edu.in

1. Introduction

Web scraping is a process where we get information from websites such as www. We get information from websites using this method, but sometimes it becomes illegal. So only that person uses the process to download the website with a license. Net scraping of presidential websites is illegal. We cannot web scrape government websites. Format, HTTP programming, HTML parsing, DOM parsing, etc. We have some data extraction methods such as We use web scraping to extract the data. We use web scraping to extract large amounts of information from websites into Excel files or databases.

We all know that health is wealth, which means that health is very important to life. Last year also proved that Covid-19 is not the only disease of concern. India has also seen cases of scarlet fever, tomato flu, black fungus, and measles in cattle, as well as nodular skin disease. Thus, we can measure viruses

with the help of web scraping. We extract information from the websites of online pharmaceutical companies so that we can understand the types of diseases in our country. Pharmaceutical companies store information about drugs in medical records.

2. Related Work

In an article called Web Scrapping published in 2017, author Bo Zhao [1] said the nice soup technique for web scraping is limited, according to this theory, if the web scraper sends too much request data, it serves as the equivalent of a denial of service. A 2012 article by Internet advertising writer Anand V. Saurkar, Kedar G. Pathare, and Shweta A Gode titled "Web Scrapping Using Collaborative-Based Networking" used HTML parsing methods, but with important details [2].

The test is included to calculate the accuracy of this research paper and the effectiveness of the process indicated.

German author JAN KINE A N D JANNA AXENBECK published in 2019 in the article Web Mining of Firm Websites: A Foundation for Web Scraping and Research Analysis, the ARGUS web scraper method used, tabs in future research, analysis of the download information network and other subsystems of the innovation ecosystem. Inclusion [3].

In a 2019 article by authors Alex Bradley and Richard J. E. James using R from Web Scraping, they use the R language web scraper method, in this article they provide a description of the R environment and how to download and install the R package [4]. Alex Bradley and Richard J. E. James' book was used in 2018 for copy and paste, parsing HTML, parsing DOM (Data Model), computer vision web analyzer, and web scraping software for web scraping method, in this article they just discuss the basics of web mining [4].

Web Scraping Using Machine Learning Author VICTOR CARLE In an article published in 2022, Web Scraping Software web scraper method using machine language, according to this theory, there are some problems in calculating truth that can increase or decrease truth [5].

S.N	Paper Title	Author	Year	Method	Limitations
9	An Overview Of Web Scraping Techniques And Tools	Rishabh Singh Tomar [2]	2018	Manual Copy and Paste, HTML Parsing, DOM (Document Object Model)	Computer vision webpage analyzers, Web Scraping Software
10	Web scraping Or Web Crawling: State Of Art, Techniques, Approaches, And Application	Anand Saurkar, Kedar Pathare, Shweta Gode [2]	2021	Parsing, BeautifulSoup, Soup, lxml	According to this view, the main issue that needs to be addressed immediately is the legality of web scraping.
11	Web Scraping Using R	Moaiaad Ahmad Khnder [4]	2019	R language	This article provides a basic introduction to the R environment and how to download and install R packages.
12	Web scraping For hospitality Research: Overview, Opportunities And Implications	Alex Bradley and Richard J. E. James [10]	2021	Python	Research in this area aims to understand the general makeup of consumers concerned and to find preliminary evidence that warrants further investigation
13	Social Media Web Scraping Using Social Media	Saram Han1,2 and Christopher K. Anderson [11]	2019	API & Regex	These studies can be further developed in practice. Developers API and Regex

Table 1. Related Work on Web Scraping

S.N	Paper Title	Author	Year	Method	Limitations
1	Web craping	Bo Zhao [1]	2017	Beautiful Soup	if the web crawler sends too many requests, this is equivalent to a denial of service according to theory.
2	A comparative Study On Web scraping	De S Sirisuriya [6]	2015	Beautiful Soup, lxml	According to this theory, most web crawlers are generally generic and most are designed to perform simple tasks.
3	Implementati on Of Web Scraping For E-Commerce Website	Prof. P.S. Gaikwad, Kaushal Parmar, Rohit Yadav, Datta Supekar [7]	2021	Beautiful Soup, Selenium, MySQL-connector	Try to improve the accuracy of product content.
4	A Study On Web Scraping	Niranjan Krishna, Anvith Nayak, Sana Badagan [8]	2022	Python community and library resources	According to this view, web scrapers and the Unauthorized Real Estate Law No. As you may be liable for damages
5	Web scraping	Vojtech Draxl [9]	2018	Manual Scraping,	According to this theory, personal data can only be collected with the consent of the affected person. Data Extraction from websites HTML Parsing, DOM Parsing, XPath, APIs
6	Web scraping Using machine Learning [5]	ENGINEERING [5]	2022	Machine Learning	According to this theory, there are some problems in calculating the truth that can increase or decrease the accuracy.
7	Web Mining Of Firm Websites: A	Felix Speckmann [3]	2019	Framework for Web Scraping	ARGUSwebscraperAnalyze the knowledge download network and include other subsystems of the innovation ecosystem in future research.
8	A Study On Web scraping	Eloisa Vargiu, Mirko Urru1 [8]	2019	Manual Copy and Paste, HTML Parsing, DOM (Document Object Model)	Depending on the main purpose, different areas can be disposed of.

3. Methodology Used

When we send a request to a server, the server sends the requested information or website in response and displays it in our browser. We convert HTML files to Python files before displaying the browser server response. We also save website data in CSV files. The following steps are used to accomplish the task:

3.1. Fetching the HTML content (Using Requests).

Our first step is to get the HTML content using the HTML request as a string. With the help of the application library, we can fetch the file using the given URL. Here we collect our data in different ways such as receiving, publishing, and deletion.

3.2. Parse the HTML (Using HTML Parse & BeautifulSoup)

HTML Parser is used to convert HTML string to format. Where we see the tree model in the law. HTML Parser is used to convert HTML string to format.

3.3. HTML Tree traversal.

We see the tree model in the law. Image-Display DOM tree-like model. With the help of such a tree, we get information from web pages. We use the class and id given for the current

element in the HTML code of the web page from the data we dump. Nice soup is a pattern in Python for retrieving data in

a hierarchical fashion. Retrieve, Share, and Delete modes to process our data.

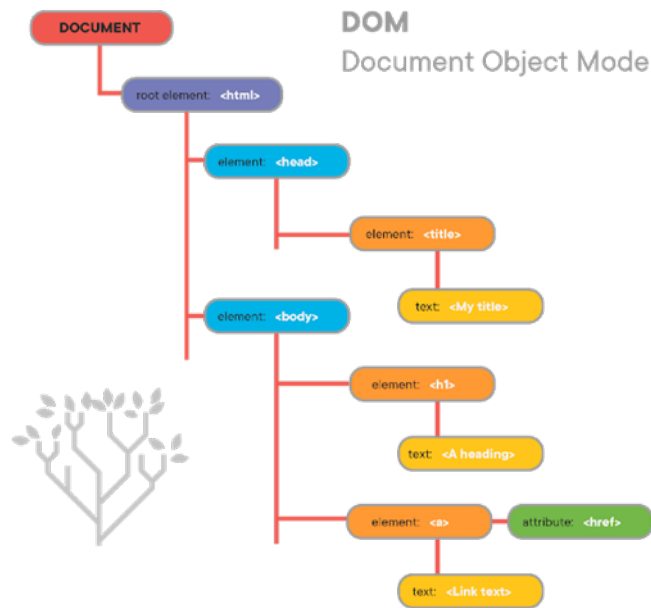


Figure 1. DOM tree like structure

3.4. Keep the information in a CSV report.

With the help of this tree, we extract information from web pages. We use the class and id given for the current element in the HTML code of the web page from the data we dump.

4. Results

As we can see from Figure 2, there are columns for drug names, manufacturers, drug types, and prices. This is the information we get from a website called Net meds, a pharmacy site. With the help of this information, we learn which drugs are most needed, and then we find the most common diseases today.

	A	B	C	D	E
1	med_name	med_manuf	med_gen	med_price	
2	DOL 3A Tablet 10'S	3A Pharmaceuticals	DICLOFEN	Rs.55.00	
3	Bitatel 40mg Tablet 10'S	4 Care Life Sciences	Telmisart	Rs.50.00	
4	Creolip 5mg Tablet 10'S	4 Care Life Sciences	Rosuvasta	Rs.45.00	
5	Creolip 20mg Tablet 10'S	4 Care Life Sciences	Rosuvasta	Rs.143.00	
6	Creolip F 10/160mg Tablet 10'S	4 Care Life Sciences	FENOFIBR	Rs.143.00	
7	Creolip Gold 10mg Tablet 10'S	4 Care Life Sciences	Rosuvasta	Rs.140.00	
8	Enzycare 10mg Tablet 10'S	4 Care Life Sciences	Escitalopr	Rs.60.00	
9	Enzycare Plus 10mg Tablet 10'S	4 Care Life Sciences	CLONAZEF	Rs.81.00	
10	Enzycare LC Tablet 10'S	4 Care Life Sciences	CLONAZEF	Rs.82.00	
11	Enzycare 5mg Tablet 10'S	5 Care Life Sciences	Escitalopr	Rs.38.50	
12	Andlazide M Tablet 10'S	A N Pharmaceuticals Pvt Ltd	GUCLAZID	Rs.94.00	
13	Anoflox O Tablet 10'S	A N Pharmaceuticals Pvt Ltd	Ofloxacin	Rs.77.00	
14	Clonax 0.5mg Tablet 10'S	A N Pharmaceuticals Pvt Ltd	CLONAZEF	Rs.29.19	
15	Clonax 2mg Tablet 10'S	A N Pharmaceuticals Pvt Ltd	Clonazep	Rs.61.71	
16	Clonax 0.25mg Tablet 10'S	A N Pharmaceuticals Pvt Ltd	CLONAZEF	Rs.18.25	
17	Clonax 1mg Tablet 10'S	A N Pharmaceuticals Pvt Ltd	CLONAZEF	Rs.41.29	
18	D Cuf Syrup 100ml	A N Pharmaceuticals Pvt Ltd	CHLORPHI	Rs.58.00	
19	Inlosa 10'S	A N Pharmaceuticals Pvt Ltd	Otc	Rs.120.00	
20	Nyme Plus Tablet 10'S	A N Pharmaceuticals Pvt Ltd	NIMESULI	Rs.30.00	
21	ULTRANURION PLUS CAPSULE 10'S	A N Pharmaceuticals Pvt Ltd	CAROTEN	Rs.341.60	
22	CPZ 50mg Tablet 10'S	A N Pharmacia Laboratories Pvt Ltd	CHLORPR	Rs.4.85	
23	Niclonz 4mg Pastille 10'S	A S V Laboratories India Pvt Ltd	Nicotine 4	Rs.88.00	
24	Quitsure 4mg Tablet 10'S	A S V Laboratories India Pvt Ltd	Nicotine 4	Rs.39.80	
25	Tazcan 4.5gm Injection	A S V Laboratories India Pvt Ltd	Piperacil	Rs.461.72	
26	Exito Plus Tablet 10'S	AAREEN HEALTHCARE	CLONAZEF	Rs.74.90	
27	Rifago 550mg Tablet 10'S	AAREEN HEALTHCARE	RIFAXIMIN	Rs.298.00	
28	Topfenac SP Tablet 10'S	AAREEN HEALTHCARE	ACECLOFE	Rs.79.90	
29	Aof 200Mg Tablet 10'S	Aabtac Lifecare Pvt Ltd	OFLOXACI	Rs.50.00	
30	Aof Estra 400Mg Tablet 10'S	Aabtac Lifecare Pvt Ltd	OFLOXACI	Rs.98.00	

572	Finecef 500mg Injection 1'S	Abbott Health Care Pvt.Ltd	Ceftriaxone 500mg
573	Finecef 250mg Injection	Abbott Health Care Pvt.Ltd	Ceftriaxone 250mg
574	Finecef CV 200/125mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 200mg+Clavulanic Acid 125mg
575	Finecef CV 50/31.25mg Dry Syrup 30ml	Abbott Health Care Pvt.Ltd	Cefpodoxime 50mg+Clavulanic Acid 31.25mg
576	Finecef CV 200/125mg Tablet 6'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 200mg+Clavulanic Acid 125mg
577	Finecef CV 100/62.5mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 100mg+Clavulanic Acid 62.5mg
578	Finecef CV DT 100/62.5mg Tablet 6'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 100mg+Clavulanic Acid 62.5mg
579	Finecef DT 50mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 50mg
580	Finecef LB 50/60mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 50mg+Lactobacillus 60MMU
581	Finecef O 200mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Cefpodoxime 200mg+Ofloxacin 200mg
582	Finecef T 500/62.5mg Injection 1ml	Abbott Health Care Pvt.Ltd	Ceftriaxone 500mg+Tazobactam 62.5mg
583	Finecef T 250/31.25mg Injection 1'S	Abbott Health Care Pvt.Ltd	Ceftriaxone 250mg+Tazobactam 31.25mg
584	Finecef T 1000/125mg Injection 1ml	Abbott Health Care Pvt.Ltd	Ceftriaxone 1000mg+Tazobactam 125mg
585	Flageryl 1000mg Tablet 2'S	Abbott Health Care Pvt.Ltd	Secnidazole 1000mg
586	Flagyl Infusion 100ml	Abbott Health Care Pvt.Ltd	METRONIDAZOLE 500MG+500U/ml CHLORIDE 800MG
587	Flagyl Suspension 100ml	Abbott Health Care Pvt.Ltd	Metronidazole 200mg
588	Flagyl Suspension 30ml	Abbott Health Care Pvt.Ltd	Metronidazole 200mg
589	Flagyl F 30/100mg Suspension 60ml	Abbott Health Care Pvt.Ltd	Furazolidone 30mg+Metronidazole 100mg
590	Flagyl F 100/400mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Furazolidone 100mg+Metronidazole 400mg
591	Flavid 200mg Tablet 10'S	Abbott Health Care Pvt.Ltd	Ofloxacin 200mg
592	Flavid 100/100mg Suspension 30ml	Abbott Health Care Pvt.Ltd	Norfloxacin 100mg+Trivudazole 100mg
593	Flavip 100mg Infusion 100ml	Abbott Health Care Pvt.Ltd	CIPROFLOXACIN 100MG
594	Flavip 250mg Tablet 10'S	Abbott Health Care Pvt.Ltd	CIPROFLOXACIN 250MG
595	Flavip 500mg Tablet 10'S	Abbott Health Care Pvt.Ltd	CIPROFLOXACIN 500MG
596	Flavip T2 Tablet 10'S	Abbott Health Care Pvt.Ltd	CIPROFLOXACIN 500MG+TINIDAZOLE 600MG
597	Flubet 150mg Tablet 1'S	Abbott Health Care Pvt.Ltd	FLUCONAZOLE 150MG
598	Fluothane Liquid 250ml	Abbott Health Care Pvt.Ltd	Halothane 0.74%
599	Folliderm Tablet 10'S	Abbott Health Care Pvt.Ltd	Biotin+Zinc+Manganese+Selenium+Zinc
600	Follihair A Tablet 10'S	Abbott Health Care Pvt.Ltd	BIOTIN 30MG+50VIA ISOLAVONES 50MG+SHACCHAROMYCES 100MG+ELEMENTAL IRON 80MG

Figure 2. DOM tree like structure

4. Conclusion

Web scraping is a process where we extract data from the web, convert the HTML code into a Python file, and then save it as a CSV file. Web Scraping Application in Medicine to collect data on the most used drugs to find the most common diseases by drug type. a website called Net meds, a pharmacy site. With the help of this information.

References

- [1] B. Zhao, "Web Scraping," *Encyclopedia of Big Data*, pp. 1–3, 2017, doi: 10.1007/978-3-319-32001-4_483-1.
- [2] A. V. Saurkar, K. G. Pathare, and S. A. Gode, "An Overview on Web Scraping Techniques and Tools," 2018.
- [3] J. Kinne and J. Axenbeck, "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," *SSRN Electronic Journal*, Sep. 2018, doi: 10.2139/SSRN.3240470.
- [4] A. Bradley and R. J. E. James, "Web Scraping Using R," *Adv Methods Pract Psychol Sci*, vol. 2, no. 3, pp. 264–270, Sep. 2019, doi: 10.1177/2515245919859535/SUPPL_FILE/BRA DLEY_AMPPSOPENPRACTICESDISCLOSUR E-V1-0.PDF.
- [5] I. T. F. O. T. D. P. I. A. C. T. A. T. M. F. O. S. C. S. A. Engineering and S. Cycle, "Web Scraping using Machine Learning," 2020, Accessed: Oct. 15, 2023. [Online]. Available: <https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-281344>
- [6] S. D. S. Sirisuriya, "A Comparative Study on Web Scraping," 2015.
- [7] K. Parmar, R. Yadav, and D. Supekar Associate Professor, "Issue 6 www.jetir.org (ISSN-2349-5162)," *JETIR2106682 Journal of Emerging Technologies and Innovative Research*, vol. 8, 2021, Accessed: Oct. 15, 2023. [Online]. Available: www.jetir.org
- [8] N. Krishna, A. Nayak, S. Badagan, and C. Jetty, "A study on Web Scraping," *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, vol. 9, pp. 2394–2320, 2022, Accessed: Oct. 15, 2023. [Online]. Available: <https://www.researchgate.net/profile/Bo-Zhao-3/publication/317>
- [9] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering- based approach to web advertising," *Artif. Intell. Res.*, vol. 2, no. 1, Nov. 2012, doi: 10.5430/AIR.V2N1P44.
- [10] S. Han and C. K. Anderson, "Web Scraping for Hospitality Research: Overview, Opportunities, and Implications," *Cornell Hospitality Quarterly*, vol. 62, no. 1, pp. 89–104, Feb. 2021, doi: 10.1177/1938965520973587/ASSET/IMAGES/LARGE/10.1177_1938965520973587-FIG9.JPEG.
- [11] L. C. Dewi, Meiliana, and A. Chandra, "Social Media Web Scraping using Social Media Developers API and Regex," *Procedia Comput Sci*, vol. 157, pp. 444–449, Jan. 2019, doi: 10.1016/J.PROCS.2019.08.237.