Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Review Maker: A Proposed Web Scraping Tool for Predicting Health Score of Company

[1]Dr. Ashutosh Sharma, [2]Dr. Nikita Joshi Mishra, [3]Ms. Archana Gupta, [4]Dr. Vikas Goel,

[1]Associate Professor IMS Ghaziabad, (University Courses Campus)

[2]Assistant Professor IMS Ghaziabad, (University Courses Campus)

[3]Assistant Professor IMS Ghaziabad ,(University Courses Campus)

[4]ProfessorKiet Group of InstitutionsDelhi-NCR, Meerut Road,Ghaziabad (Uttar Pradesh)

Abstract—The issue of today's technological world is that there are various websites for the same product. Customers find it difficult to choose between those websites and sometimes they even get wrong reviews too. In and around the e-commerce industry, scraping product reviews from e-commerce websites has become one of the most important competitive intelligence operations.In the private sector, web scraping: the extraction of patterned data from web pages on the internet, was invented for achieving corporate goals.But it has significant advantages to people looking for company health score ratings. In this paper, a scraping review application: Review maker has been proposed and implemented using MERN (MongoDB, Express, React, Node).Then the sentimental analysis is done over the collected tweets having the company's tag. The proposed applicationis based on an asynchronous function. That works together to collect the review from different sites and tweets from Twitter to create a generalized health score. Prediction is based on the information provided by the user and the dataset is based on the factors that are responsible for the health condition of the company.

Index Terms—*web scraping, MERN, sentimental analysis, reviews, puppeteer.js*

Introduction

In and around the e-commerce industry, scraping product reviews from e-commerce websites has become one of the most important competitive intelligence operations. Product reviews on e-commerce websites are a fantastic way to get honest feedback from real customers. This implies that if you're a manufacturer looking for in-depth information about your product, e-commerce product review pages are the place to go. While product evaluations may be found on a variety of e-commerce sites, not every firm has the knowledge, technology, or resources to crawl and extract reviews from e-commerce sites in an automated manner. [1]

eCommerce is sweeping the globe, and there are plenty of product reviews to be found on these sites. The major benefit for manufacturing firms is that these reviews are objective, which will help them better understand their customers and serve them. Here are some of the most common uses for scraping product reviews. [2]

- Understand your consumer preferences
- Brand monitoring
- Competitor analysis
- Natural language processing
- Fraud Detection

Web scraping, also known as web harvesting or web data extraction, is a technique used to extract data from websites on the Internet. It is often used for a variety of purposes, including data analysis, research, and automation. The practice of web scraping has been around for many years and was originally used by researchers and academics to collect data from websites for analysis and study. As the internet became more commercialized, businesses began to see the value in web scraping for competitive intelligence and market research. Today, many companies use web scraping to gather data about their competitors, track pricing and product information, and monitor customer reviews and sentiment.

Web Scraping is a great technique of extracting unstructured data from the websites and transforming that data into structured data that can be stored and analysed in a database. Web Scraping is also known as web data extraction, web data scraping, web harvesting or screen scraping. Web scraping is a form of data mining. The overall goal of the web scraping process is to extract information from a websites and transform it into an understandable structure like spreadsheets, database or a comma-separated values (CSV) file as shown in Figure 1 [3].

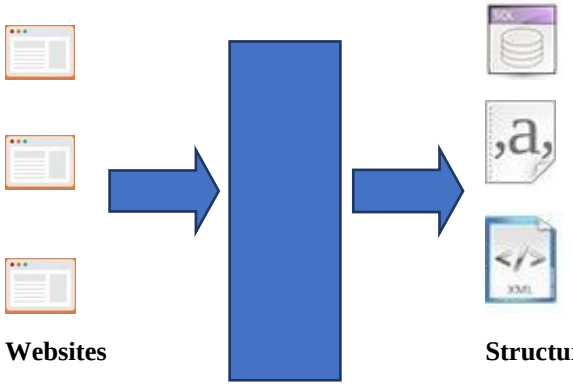

Fig. 1. Basic Architecture of Web Scraping

Data is very important for businesses and organizations as it assists their decision making and especially, currently, most of the data can be found on the internet [19]. Data acquisition is the first phase in any data science study and development; it is the step in which the data obtained from private sources such as firm sales records and financial reports, or from public sources such as journals, websites, and open data, or by purchasing data [20]. Website analysis, website crawling, and data organizing are the three primary, interwoven processes of online scraping [21]. Web scraping differs from data mining in that the latter entails data analysis, whereas collecting data is immaterial in this situation. Data mining also necessitates the use of sophisticated statistical techniques [22] Due to the wide number of accessible tools and libraries that offer efficient implementations of much of the required functionality, web scraping is a pretty simple process in general. The ability to send custom HTTP requests with different headers and payloads is standard feature of most web scraping programs [23].

Web Scraping Software are the tools that are used to automate the manual copy paste work to gather large amount of data from websites like directory sites, real estate sites, classified websites and job boards. Suppose you want to scrape real estate property details of any city then you need to appoint few guys to copy and paste details from various websites to excel by visiting each property page. This way it will take days and even months to get your property data ready to use. So web scraping can automate the manual work programmatically by visiting each page and extract data from pages and parsing the html pages. There are number of Web Scraping Software that available in market that can help you to scrape data from any website you want.[24]

Customer reviews can make or break an online reputation. Unfortunately, most customers that are motivated to leave reviews often end in a negative review Maker is designed specifically to net you more positive customer reviews spread across any review site choose. The purpose of this research paper is:

- To design a review maker that may collect reviews from different websites for a particular company.
- To usethe data from many reliable websites and
- To provides reviews in nutshell.

The main objectives of the proposed framework are:

- Provide reliable data to the customers.
- Summarize the review in a number.

I.RELATED WORK

The authors presented empirical research to learn how developers utilize distinct language features in JavaScript. He intendedto apply this information to help future JavaScript extensions. The authors examined over one million unique scripts (over 80 MLOC) from a variety of sources. It includes JavaScript programs found in the wild by a spider.These are better JavaScript programs found in the top 100 Alexa URLs. JavaScript programs with new language features found in Firefox Add-ons.These are widely used JavaScript libraries and Node.js applications. The proposed research reveals a widespread misunderstanding of recently added JavaScript capabilities.[7]

Web development technology has come a long way. The addition of JavaScript has significantly improved the client's interactivity. To solve these challenges, this work proposes a static detection strategy based on semantic analysis. The method generates an abstract syntax tree from JavaScript source code before automatically transforming it to syntactic unit sequences. The FastText technique is used to train word vectors. [8]

The authors presented a design framework for creating reactive systems like client-side web apps and smartphone customization apps. These were assemblages of independent cross-cutting situations that were knitted together during runtime. These included a single-threaded implementation of behavioral programming (BP) in JavaScript and coupled visual programming with Google Blockly. [9]

Hospitality experts are becoming increasingly interested in studying the internet travel sector. Even though many academics have sought to better understand the internet travel industry by using analytical models, experiments, or survey data.These studies typically fall short of capturing the full complexity of the industry. The authors sought to provide tools and approaches for hospitality researchers in this effort. Consumers utilize readily available data to make travel decisions, and this study augmented their typical data sources with readily available data. In the essay, the authors suggested how to effectively collect/scrape publicly available internet hotel data. The writers gathered information from the internet from several places, including online travel companies, review sites, and hotel brand websites.The authors explored some of the exciting potentials for using various data sources, as well as some of the constraints that must be addressed when evaluating internet data.[10]

This article examines the many techniques for doing political analysis utilizing numerous online sources of information. Social networking has a considerable influence on the political field in some cultures, and numerous techniques have been utilized to examine various political elements and tactics to be pursued. The goal of this article is to have a better understanding of these techniques so that potential voters may make educated judgments. The essential vocabulary for web scraping is covered first, followed by several instances of political analysis initiatives that have employed web scraping.[11]

The tools and libraries used in this study are classified into three categories: partial tools, libraries and frameworks, and complete tools. Over the previous few years, these tools and libraries have been built. They're commonly employed to gather information and turn it into structured data for text-processing applications. The terms "web scraping" and "web crawling" are defined in this article. The authors also identified the available tools and showed the reader how to use one of them to develop their Web scraper. The essay finishes with a discussion on Web scraping's legality.[12]

The identification and classification of people's views, attitudes, and feelings conveyed in original statements is the goal of sentiment analysis. The evolution of social media, various critiques, forum debates, blogging, and social network activity may all be classified in many ways. On the platform, users generate large amounts of sentiment data in the form of tweets and status updates. Market research and product development firms can benefit from sentiment analysis of this data. They are increasingly relying on the public opinion expressed in various mediums to make decisions. In this work, the authors offered a method for assessing emotion or opinion in a timely and effective manner. To categorize sentiment analysis for tweets, the authors provided an approach that focuses on multilingual data analysis.[13]

Proposed Work

Web scraping software is a set of tools for gathering huge amounts of data from websites such as directories, real estate websites, classified websites, and job boards by automating the tedious copy-paste process. If you want to scrape real estate property facts in the United Kingdom, you'll need to hire a few people to visit each property page and copy and paste information from websites to excel. It will take days, if not months, to have your property data ready to utilise this manner. As a result, web scraping can programmatically automate the arduous process of visiting each page and extracting data from them, as well as parsing the html pages. There is a variety of Web Scraping Software available on the market that may assist you in scraping data from any website.

Data scraping is the process of extracting data from an electronic file using computer software. Web scraping is the technique of utilizing software to retrieve data from HTML files on the internet. This information is frequently presented in the form of patterned data, such as tables or lists. Application programming interfaces (APIs) are sets of commands used by programs to interact with websites and retrieve data (APIs). These APIs may be trained to extract patterned data from individual web pages or all comparable pages on a website. Alternatively, APIs may be used to automate website activities, such as clicking links inside a page and extracting data.

The application may generate a company’s health score by scrapping reviews from different sites like indeed, kununu, etc using puppeteer.js; and the sentimental analysis done over the recent tweets with the company’s tag, the technologies used may be web scrapping and JavaScript development. By this proposed framework, an online review portal may be built where users may find the review of any company in a form of a general health score. The review maker may collect the reviews from different websites for a particular company. It uses data from many reliable websites and provides reviews in a nutshell. A system based on an asynchronous function works together to collect the review from different sites and tweets from Twitter to create a generalized health score. The dataset is based on the elements that are responsible for the company's health status, and the prediction is based on the information given by the user.

Generate a company’s health score by scrapping reviews from different sites:

- The application may generate a company’s health score
- The sentimental analysis done over the recent tweets
- The technologies used may be web scrapping, MERN stackdevelopment
- The proposed framework: REVIEW MAKER.

Steps for the execution of the proposed framework: REVIEW MAKER

1. The reviewed library runs through being called by the routes of the application in which it is being integrated. The different routes are called for different sites. Figure2 depicts the sample code to fetch data from Comparably in the form of a numeric review.

Fig.2. The code to fetch data from Comparably in the form of a numeric review

```
fetchkununu = async (name,callback)=>{
    var url = 'https://www.kununu.com/de/' + name;
    const browser = await puppeteer.launch({headless:true})
    const page = await browser.newPage();
    await page.goto(url,{waitUntil:'domcontentloaded'});
    await page.addScriptTag({url:'https://code.jquery.com/jquery-3.2.1.min.js'});
    const resut = await page.evaluate((data)=>{
        var obj = $( '.index__value__1EG5h.h2.h3-semibold-tablet').text()
        data.push(parseFloat(obj));
        return parseFloat(obj);
    },data)
    console.log(resut);
    await page.close();
    await browser.close();
    callback(resut,true);
}
```

2. The function called for different functions in the library is done through the callback function of javascript. The different parameters being passed in the calling function includes the company name and a returning console.log statement which is called when the function successfully returns the computed review value.

3. The calculation of the average health is based on the reviews generated from different sites which are returned as a numeric value. These numeric values are computed to calculate the average of the reviews. The resultant is then subtracted from the base value to give the health score and the final premium is obtained from this health score over the base premium.

4. Data fetched from kununu, the reviewing site is taken from the document object model through cascading style sheet selector and is opened through the headless property of the puppeteer that allows the task of fetching the review in the background. The sample code is depicted in figure 3.

```
fetchcomparably = async (name,callback)=>{
    var url = 'https://www.comparably.com/companies/'+name
    const browser = await puppeteer.launch({headless:true})
    const page = await browser.newPage();
    await page.goto(url,{waitUntil:'domcontentloaded'});
    await page.addScriptTag({url:'https://code.jquery.com/jquery-3.2.1.min.js'});
    const resut = await page.evaluate((data)=>{
        var obj =    $('.numerator').text()
        data.push(parseFloat(obj));
        return parseFloat(obj);
    },data)
    console.log(resut);
    await page.close();
    await browser.close();
    callback(resut,true);

}
```

Fig. 3. The code to fetch data from Kununu in the form of a numeric review

5. The landing page of the site takes the different kinds of company data as input from the form and passes that data for generating the health score and premium calculation. The send keyword of express does the work of sending JSON responses from a route to another. This route for the landing page of the site, calls for the functions to generate the review based on the data scraped from different sites. The sample function is depicted in figure 4.

```
app.get("/insurace",(req,res)=>{
    if(!req.query.dis){
        res.send({
            error:"please provide dis"
        })
    }

    var compData = {
        regNumValue : req.query.regNum,
        locationValue : req.query.location,
        companyValue : req.query.company
    }

    fs.writeFile("compData.json",JSON.stringify(compData), function (err) {
        if (err) throw err;
        console.log('Saved!');})
    var pre = premium(10,10,10,10);

    //taking the premium amount in the format of
    // premium(valueOfCargo,totalDistance,typeOfVehicle ,typeOfPackageCost)
    //var pre = premium(1000000,3500,"low" ,10);
    var pre = premium(req.query.CargoValue,req.query.dis,req.query.vehicle ,10);
    var final =  finalPremium(pre, health);
   res.send({
            CargoValue:req.query.CargoValue,
            dis:req.query.dis,
            vehicle:req.query.vehicle,
            package:req.query.package,
            finalPre : final
    })
    console.log("company details"+ req.query.CargoValue+' '+req.query.dis+' '+req.query.vehicle+' '+req.query.package)

})
```

Fig. 4. This route for the landing page of the site, calls for the functions to generate the review based on the data scraped from different sites

6. The reviews and results generated from the reviewing sites are reflected in the frontend although querying the data in the document object model of the resulting site. This is done by passing the parameter results in the query as a set of key-value pairs.

7. The data is received as a form of JSON API response from one route to another with the help of the get function of express.js and the health score is calculated at the data receiving route.

Analysis

Following are the technologies used to achieve the result of the proposed model for collecting the review from different sites and tweets from Twitter to create a generalized health score. Prediction is based on the information provided by the user and the dataset which is based on the factors that are responsible for the health condition of the company.

- Handlebars, CSS [14]
- JavaScript [15]
- Third PARTY APIs [16]
- Twitter API [17]
- Puppeteer.js, Node.js, Express.js [18]
- Web Designing Tools [18]

The calculation of the average health is based on the reviews generated from different sites which are returned as a numeric value. These numeric values are computed to calculate the average of the reviews. The different reviews of the company are represented in Table 1.Company reviews show the comparison of different

companies from different sites. These are the values that are fetched by the application that helps to calculate the health score.

Table 1. Comparison of different company's reviews from different sites

Company	Indeed	Inhersight	Kununu	Comparibly
Google	4.3	3.8	4	4.7
Facebook	4.1	3.5	4.3	4.7
ZS Associates	3.5	3.4	4.4	4.3
Ninjacart	4	4	4.7	3.8
Jio	3.9	3.7	3.8	3.7
Honeywell	3.7	2.8	3.1	2.4
Bajaj Finance	3.9	2.3	3	2.2
FIS	3.5	2.9	2.6	3
Hewlett Packard	3.8	3.7	3.4	3.6

The premium calculation page of the site takes the different kinds of company data as input from the form and passes that data for generating the premium that depends on the varying cost based on the health cost as shown in Table 2. Premium Comparison Based on Details describes variation in the premium calculation of different companies based on the health score, the variation described serves as the base of discount on the base cost.

Table 2: Comparison of Premium Based On various parameters

Company	Value of Cargo(INR)	Distance(Km)	Delicacy type	Health Score	Base	varying cost	Premium (INR)
Google	1000000	3500	Low Risk	8.2	58150	4768.3	53381
Facebook	1000000	3500	Low Risk	8.3	58150	4826.2	53323
ZS Associates	1000000	3500	Low Risk	7.8	58150	4535.7	53614
Ninjacart	1000000	3500	Low Risk	8.25	58150	4797.3	53352
Jio	1000000	3500	Low Risk	7.55	58150	4390.3	53759
Honeywell	1000000	3500	Low Risk	6	58150	3489	54661
Bajaj Finance	1000000	3500	Low Risk	5.7	58150	3314.5	54835
FIS	1000000	3500	Low Risk	6	58150	3489	54661
Hewlett Packard	1000000	3500	Low Risk	7.25	58150	4215.8	53934

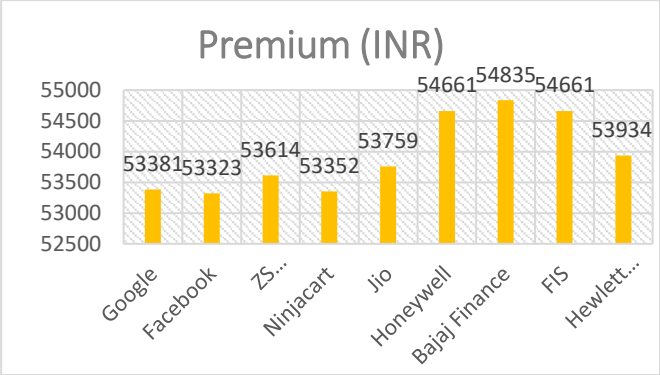

Fig.5. Graphical representation of Premium (in INR) of different companies

In figure 5, the graphical representation of premiums of different companies based on the parameters has depicted the value of the cargo, distance, delicacy type, health score, and variation from the base.

Table 3.Health Score shows the values derived from the library based on the data scraped from different scraping sites

Company	Health Score
Google	8.2
Facebook	8.3
ZS Associates	7.8
Ninjacart	8.25
Jio	7.55
Honeywell	6
Bajaj Finance	5.7
FIS	6
Hewlett Packard	7.25

In figure6, the calculated health score of different companies is based on the reviews collected from different sites are analyzed. This health score helps in calculating the premium of the company.

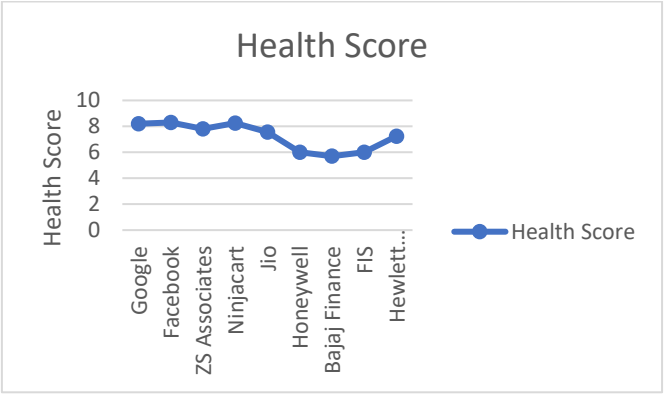

Fig.6. Graphical representation of health score of different companies

Figure 6 depicts the graphical representation of health scores of different companies.From table 3 and figure 6, it is observed that the health score of facebook is maximum and the health score of Honeywell and FIS is minimum.

Conclusion

Develop an integrable JavaScript library that generates the review of a company in the form of a health score; the result thus generated should be fully reliable and fast derived. The JavaScript integrable library is ready and generate correct results in the form of a health scare in the stipulated time frame.The Review Maker application may generate a company's health score by scrapping reviews from different sites like indeed, kununu, etc. using MERN (MongoDB, Express, React, Node) that make up the stack. The sentimental analysis is done over the recent tweets with the company's tag.The technologies used may be web scrapping and JavaScript development. The proposed Review Maker application collects the reviews from different websites for a particular company. It uses data from many reliable websites and provides reviews in a nutshell.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Vikas Goel conduct the research. Narendra kumar, Raju Ranjan and Amit Kumar Gupta collected and analyzed the data and write the paper also. All authors had approved the final version.

References

[1] G. Liu, S. Fei, Z. Yan, C. Wu and S. Tsai, "An Empirical Study on Response to Online Customer Reviews and E-Commerce Sales: From the Mobile Information System Perspective", *Mobile Information Systems*, vol. 2020, Article ID 8864764, 12 pages, 2020.

[2] Y.K.Dwivedi, E. Ismagilova, D.L. Hughes, J. Carlson, R. Filieri, J. Jacobson, V. Jain, H. Karjaluoto, H. Kefi, A.S. Krishen, and V. Kumar, "Setting the future of digital and social media marketing research: Perspectives and research propositions",*International Journal of Information Management*, vol. 59, pp.102168, 2121.

[3] A.V.Saurkar, K.G.Pathare, and S.A.Gode, "An overview on web scraping techniques and tools", *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. *4, no.* 4, pp.363-367, 2018.

[4] Use of Callback function *https://www.freecodecamp.org/news/javascript-callbackfunctions-what-are-callbacks-in-js-and-how-to-use-them*, last accessed 2021/08/09.

[5] How to use map and filter function *https://masteringjs.io/tutorials/fundamentals/map-filter*, last accessed 2021/08/09.

[6] Components in React *https://reactjs.org/docs/components-and-props.html*, last accessed 2021/08/09.

[7] M. Hafiz, S. Hasan, Z. King, and A. Wirfs-Brock, "Growing a language: An empirical study on how (and why) developers use some recently-introduced and/or recently-evolving JavaScript features", *Journal of Systems and Software*, Vol. 121, pp.191-208, 2016.

[8] Y. Fang, C. Huang, Y. Su, and Y. Qiu, "Detecting malicious JavaScript code based on semantic analysis", *Computers & Security*, vol. 93, pp.101764, 2020.

[9] Ashrov, A. Marron, G. Weiss and G. Wiener, "A use-case for behavioral programming: an architecture in JavaScript and Blockly for interactive applications with cross-cutting scenarios", *Science of Computer Programming*, vol. 98, pp.268-292, 2015.

[10] S. Hanand C.K. Anderson, "Web scraping for hospitality research: Overview, opportunities, and implications", *Cornell Hospitality Quarterly*, vol. 62, no. 1, pp.89-104, 2021.

[11] N. Varela, O.B.P. Lezamaand M. Charris, "Web Scraping and Naïve Bayes Classification for Political Analysis",*In Proceedings of International Conference on Intelligent Computing, Information and Control SystemsSpringer, Singapore*, pp. 1-8, 2021.
[12] H. Nigamand P. Biswas, "Web Scraping: From Tools to Related Legislation and Implementation Using Python",*In Innovative Data Communication Technologies and Application Springer, Singapore*,pp. 149-164, 2021.
[13] P. Goel, V. Goel and A.K. Gupta, "Multilingual data analysis to classify sentiment analysis for tweets using nlp and classification algorithm", In *Advances in Data and Information SciencesSpringer, Singapore*, pp. 271-280, 2020.
[14] A. Mardan, "Template Engines: Jade and Handlebars",*In Practical Node. JsApress, Berkeley, CA*, pp. 71-102,2014.
[15] A. Ntantogian, P. Bountakas, D. Antonaropoulos, C. Patsakis and C. Xenakis, "NodeXP: NOde. js server-side JavaScript injection vulnerability DEtection and eXPloitation", *Journal of Information Security and Applications*, vol. *58*, pp.102752, 2021.
[16] What is API? *https://www.mulesoft.com/resources/api/what-is-an-api*, last accessed 2021/08/09.
[17] Twitter API *https://developer.twitter.com/en/docs*, last accessed 2021/08/09.
[18] F. Kaimer and P. Brune, "Return of the js: Towards a node. js-based software architecture for combined cms/crm applications", *Procedia Computer Science*, vol. 141, pp.454-459, 2018.
[19] Almaqbali, I. S., Al Khufairi, F. M., Khan, M. S., Bhat, A. Z., Ahmed, I. (2019). Web Scrapping: Data Extraction from Websites. Journal of Student Research.
[20] Chaulagain, R. S., Pandey, S., Basnet, S. R., Shakya, S. (2017). Cloud based web scraping for big data applications. 2017 IEEE International Conference on Smart Cloud (SmartCloud), (pp. 138–143).
[21] Milev, P. (2017). Conceptual approach for development of web scraping application for tracking information. Economic Alternatives, (3), 475-485.
[22] Krotov, V., Silva, L., 2018. Legality and ethics of web scraping, Twentyfourth Americas Conference on Information Systems, New Orleans.
[23] Yannikos, Y., Heeger, J., Brockmeyer, M. (2019). An Analysis Framework for Product Prices and Supplies in Darknet Marketplaces. Proceedings of the 14th International Conference on Availability, Reliability and Security. New York, NY, USA: Association for Computing Machinery.
[24] Sirisuriya, De S. "A comparative study on web scraping." (2015).

Authors

Dr.Ashutosh Sharma is working as Associate Professor & Head-(Student welfare & Examination) at IMS Ghaziabad, University Courses Campus. He has an experience of more than two decades in the field of Academics and Administration at various levels. He did his Ph.D. (Computer Science) from SR University. He has received an award for "Best Faculty" in the year 2022. He has more than 16 years of total experience in Academics, where he has handled various roles of teaching and administrator. He is a certified (CompTIA Security+) in the field of Security .

Dr. Nikita Joshi Mishra is working as an Assistant Professor at IMS Ghaziabad, University Courses Campus. She has an experience of 11 years in academics and research. She was awarded her Ph.D. in Computer Science and Engineering from Amity University Uttar Pradesh. She completed her M.Tech (Computer Science & Engineering) from Amity University Uttar Pradesh and B.Tech from Uttar Pradesh Technical University. She has published 13 research papers in reputed national and international journals.

Archana Gupta is working as an Assistant Professor at IMS Ghaziabad, University Courses Campus. She has an experience of 15 years in academics and research. She is pursuing her Ph.D. in Computer Science and Engineering from JJTU, Rajasthan. She completed her MCA from Uttar Pradesh Technical University and M.Tech (IT). She has Qualified UGC NET (June 2012). She has published research papers in reputed national journal.

Dr.Vikas Goel received B.Tech. degree in Information Technology from VBS Purvanchal University in 2001, M.Tech. degrees in computer science & engineering from Shobhit University in 2009 and Ph.D. in computer science & engineering from Uttarakhand Technical University, Uttarakhand, 2017.He is currently an Associate Professor in the IT Department at KIET Group of Institutions, Ghaziabad affiliated from AKTU, Lucknow from January 2020 onwards. He has a vast experience of 19+ years of teaching in various good institutes. His research interests include mobile computing, wireless computing, broadcasting data in mobile devices, distributed computing, sentiment analysis, Blockchain.