Mathematical and Software Engineering, Vol. 8, No. 1-2 (2022), 19-23.
Varεpsilon Ltd, http://varepsilon.com
DOI 10.5281/zenodo.7294898

Using web scraping for real estate price analysis

Aleksandar Tsonev[1], **Valentina Dyankova**[2], **Yusuf Yusufov**[3]

[1]Dreamix LTD, Sofia, Bulgaria,
email: aleksandar.tsonev@dreamix.eu

[2]Konstantin Preslavsky University of Shumen, Shumen, Bulgaria,
email: v.dyankova@shu.bg

[3]Konstantin Preslavsky University of Shumen, Shumen, Bulgaria,
email: yusuf.yusufov969@gmail.com

Abstract

In the present paper, a survey is being made of the areas, in which the use of Web scraping is necessary for adaptation to dynamic changes, demanding processing and generalizing of large volume of data from different web sources. An exemplary implementation for Web scraping is presented for price analysis of the real estate market. The technology that gives possibility of integration of the intended activities in development of the application has been reviewed. The overall procedure in the researching, development and realization of the proposed application can be used as a good practice in adaptation of the knowledge and skills of the graduating students for the requirements, dynamics and organization of the work in a software company as well as for the professional co-operation between the academic lecturers and representatives of the software business.

Keywords: web scraping; data extraction; databases; Selenium; Java; SpringBoot; Hibernate

1. Introduction

Even with the abilities of new technology to present even more adapting complex solutions in all areas of contemporary life, more and more organizations acknowledge that the strategic use of data and the creation of a culture managed by data have a conclusive meaning for the competitive support of their activity. The increasing pace of digitalization in all areas of the social life makes a large volume of data. That makes the processes of their gathering and use more critical than ever. These characteristics determine the present significance of Web scraping, which extracts big amounts of data usually from websites.

The goal of this article is to introduce the usefulness of Web scraping by reviewing areas where it is used and to suggest one type of its application for price analysis for the real estate market. The suggested article synthesizes activities and solutions from the diploma project of one of the authors. The technologies used are described from the point of view of the adaptive integration of knowledge and skills of a bachelor’s degree graduating student towards the work organization in a software company. The other two authors are representatives of the software industry and academic society and in this sense the overall activity on the research, designing and realization of the suggested application can be used as a good practice both for adaptation of knowledge and skills of the graduating students towards the requirements and dynamics of a software company as well as professional collaboration between the academic lecturers and representatives of the software business in context of the central theme of ACM Computing Curricula 2020 for competency. Competency is reviewed as an integrated collectivity of knowledge studied in disciplines of the curriculum with their skillful implementation through specific technologies of the industrial area, motivated by a goal for execution of a task practically important for the present dynamics. The goal chosen for the graduation project is the Web scraping entering all areas of social life.

2. Web scraping

Web scraping is a process of retrieving valuable and interesting text information from web sites [1]. The goal of this process is the transformation of retrieved raw information into structured data, which are significant and plausible for assistance of a specific activity [2]. The present digitalization in all areas of social life is the basis for the use of Web scraping in all regions of the human activities. Some of them are:

Social life

The Web application such as Facebook, Twitter and Instagram provide easy functions for sharing information. Sometimes the information is not presented appropriately for the user or maybe he is troubled by its redundancy. In [3], the Web scraping method is used that can search information, combine it and provide it in a way consistent with the user requirements.

Through Web scraping analyzing of the social media posts from March 2020 till January 2021 research is done on the subjective use of loneliness in times of COVID-19 [4]. The language markers of emotions such as depression, anxiety, anger, hate, helplessness and sadness have been researched.

Tourism

Because more and more users research and buy hotel and tourism services online, at hand is the present interest towards systematization of the information available in the web space. Focusing on gathering of online data in multiple platforms, online tourism agents, web sites for reviews, and web sites for hotel brands in [5] are presented guidelines and realization of the gathering and systematization of publicly available online data for hotels with the goal of extracting ratings of satisfaction, performing of statistical analysis, assessment of correlation between price and rating on the webpage of the online tourism agent.

Machine Learning

Machine learning presumes that data is given to machines for them to learn and improve themselves. The Web is the ideal source for such data and Web scraping serves for their retrieval and providing for machine learning. The platform described in [6] that provides user interface for visualization of data, analysis, prognostication and investing recommendations is based on algorithms for machine learning for prognostication using data retrieved with Web scraping for purchase/sale on the Spanish market.

E-commerce

The e-commerce market has made a big jump in the last decade and it continues to develop, because the digital devices are being integrated in people’s lives and they change their behavior during purchase. In the e-commerce, the companies have to own flexible competitive solutions for determining the prices, presenting the products, multiplying of the potential clients, improvement of the quality, protection of the brand. Web scraping is exceptionally useful for price optimization. It gives the opportunity to follow the dynamic changes in the market prices and promotional events. Web scraping can automate the process of granting images and descriptions of products faster than people by using information from different suppliers for the same product. Potential clients can be found by analyzing the requirements of the audience of social networks. Analyzing reviews and reviews of the users can help the business to understand what is missing in their products and services and to identify how the competitors are distinguished. The online content can be identified fast by using Web scraping (e.g. fake products), which can damage the respective trademark. Web scrapers can help determine if the intellectual property protected with copyrights is used illegally.

Other areas

There are many more areas of use of Web scraping. For example, a surview is made in [7] of uses of Web scraping in Atrificial Intelligence, Data Science, Big Data, Cloud Computing and Cyber Security.

The variety of areas of use of the Web scraping made it difficult to choose a problem that would integrate modern dynamics, the interest of the graduate and his opportunities for contact with real business representative in the relevant field. The choice of price analysis of the real estate market as an area for implementation of the Web scraping is dictated by the rising in the last few years of interest towards investing in real estate, the interests of the student and the opportunities to discuss business logic with real estate companies. At the moment considerable interest towards the market of real estate is being observed. The variety of data and websites as well as the dependence of this market on many factors, make it incredibly hard for the end user to decide on the right moment for purchase/sale of a real estate. That’s why the system presented in [8] for prognostication of the price by extracting data from the network using Web scraping is a valuable helper. The users are often only in need of generalization of the price data to choose the optimal offer for them. The search and offering of real estate are distinguished with great dynamics and frequently with great diversion in the financial parameters, which makes generalizing of the price data particularly important. The manual analysis is slow and labour-intensive. Thus, is determined the interest of the authors towards development of an application that is using Web scraping to retrieve data for the prices of real estate, systematizes them and performs a summary.

3. Web scraping in real estate price analysis

The overall activity of researching, designing and realization of the application that uses Web scraping for price analysis of real estate market, is organized in the context of adaptation of academic knowledge of the student towards the visions and trends for technology in industry and is in line with the technologic time for developing of a graduation project.

The assignment for the graduation project included the following tasks; researching of the functional capabilities of the existing methodologies for Web scraping; familiarizing with the business logic in marketing with real estate; choice of technology used in the software industry for integrated solution of the previous two tasks; gaining knowledge and skills for work with the chosen technology; creation of an application for price analysis on real estate market. The task management platform Trello was chosen to manage the collaboration of the team members on assigned tasks. Trello is a tool that enables team management of the project development workflow. The use of boards, lists and maps makes it possible to both differentiate the assigned tasks and to track the progress on each of them. The version control system of choice was Git, with an emphasis on both the collaboration capabilities between different people and teams and the flexible branching model, allowing independent local branches to be created in the code. The latter makes it possible both to experiment with new ideas as well as to correct existing ones. The use of these two systems provides an opportunity to develop skills such as time management, organization, collaboration, attention to detail, responsibility, self-motivation, work ethic.

The conducted review of used web scraping methodologies against the set goals and their compatibility with using Java as a programming language limited the study to a comparative analysis between the capabilities of Jsoup [9,10] and Selenium [11,12,13]. The fact that Jsoup cannot handle dynamic content generated by JavaScript defined Selenium as the software tool to implement the application. The main advantage of Selenium used for its intended purpose is the use of cross-browser web drivers that build the DOM tree of each page and provide access to individual parts of the page through selectors. Selectors are the way to identify an HTML element on a web page. The fragment below is an example application of using a selector in a text retrieval function:

```
public static String getTextByXpath(WebDriver driver,String xPath){
     try {
         return driver.findElement(By.xpath(xPath)).getText();
     } catch (NoSuchElementException e) {
              log.info(String.format("Element with xpath [%s] don't exists", xPath));
              return “";
     } catch (NullPointerException e) {
               return "";
     }
}
```

The database technology used was Hibernate due to the fact that it ensures consistency of data structures and their types and facilitates the basic CRUD (create, read, update, delete) operations. Hibernate is an ORM (object-relational mapping) library that works with classes (entities) containing meta information about the connections and structure of the database and the tables in it. In order to reduce the amount of code, the Lombok library was used to generate the base methods of the classes (mainly getters and setters). CheckStyle was used to ensure code quality and following accepted formatting rules. The SpringBoot framework was chosen as the basis of the entire application due to its wide distribution in the industry and its ability to provide various features such as:

- scheduled jobs - used to retrieve information at specified time intervals
- web interface - used to visualize the processed data
- Spring IoC Container - used for creating separate extraction units/beans

The Maven system was used to manage the dependencies and build the entire application.

The developed application enables to track the maximum, minimum and average price of real estates by type, location and agency that offer the real estate for sale. This data would be useful both for proposing a sale price for a real estate and controlling the price at purchase.

4. Conclusion

The described process of researching and creating a web scraping application can be used to differentiate advantages that make this method popular among different industries. Some of them are: achieving significant results with acceptable parameters of the invested efforts; easy implementation and maintenance; accurate data extraction, preventing opportunities for manual processing errors. As the internet continues to grow and businesses become increasingly dependent on data, the industry in these days is forced to access the latest data in every field.

The development of such projects, on the other hand, enables the creation of useful links between industry and academic society. In many cases, the implementation of theoretically defined processes may require the use of specific technological solutions. Such projects can help the future realization of students, who in many cases do not have the knowledge and skills to advertise outside of academic settings.

References

[1] Erdinç Uzun, A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages, IEEE Access, volume 8, 2020, ISSN: 2169-3536.

[2] Jay M. Patel, Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale, Apress, 2020, ISBN 978-1484265758.

[3] Lusiana Citra Dewi, Meiliana, Alvin Chandra, Social Media Web Scraping using Social Media Developers API and Regex, Procedia Computer Science, Volume 157, 2019, Elsevier.

[4] Junk, Yoonwon; Lee, Yoon Kyung; Hahn, Sowon, Web-scraping the Expression of Loneliness during COVID-19, Proceedings of the Annual Meeting of the Cognitive Science Society, Volume 43, 2021, e-ISSN 1069-7977.

[5] Saram Han, Christopher K. Anderson, Web Scraping for Hospitality Research: Overview, Opportunities, and Implications, Cornell Hospitality Quarterly, Volume 62, Issue 1, 2020.

[6] Elena Hernández-Nieves, Álvaro Bartolomé del Canto, Pablo Chamoso-Santos, Fernando de la Prieta-Pindato, Juan M. Corchado-Rodríguez, A Machine Learning Platform for Stock Investment Recommendation Systems, Distributed Computing and Artificial Intelligence, 17th International Conference, Italy, 17th–19th June, 2020.

[7] Moaiad Ahmad Khder, Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application, International Journal of Advances in Soft Computing and its Applications, Vol. 13, No. 3, November 2021.

[8] Tomasz Jach, Web Scraping Methods Used in Predicting Real Estate Prices, Advances in Computational Collective Intelligence, 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29 – October 1, 2021.

[9] Pete Houston, Instant jsoup How-to, Packt Publishing, 2013, ISBN 978-1782167990.

[10] Ryan Mitchell, Instant Web Scraping with Java, Packt Publishing, 2013, ISBN 978-1849696883.

[11] Boni Garcia, Hands-On Selenium WebDriver with Java: A Deep Dive into the Development of End-to-End Tests, O'Reilly Media, 2022, ISBN 978-1098110000.

[12] Pallavi Sharma, Selenium with Java – A Beginner’s Guide: Web Browser Automation for Testing using Selenium with Java, BPB Publications, 2022.

[13] Kevin Sahin, Java Web Scraping Handbook, Leanpub, 2018-07-26.