Andrii Atoian
1st year Master student,
Economic Cybernetics Department,
Taras Shevchenko National University of Kyiv,
Kyiv, Ukraine
fike1@ukr.net

APPLICATION OF DYNAMIC DATA SCRAPING TECHNOLOGY FOR E-COMMERCE WEBSITES

Key words: e-commerce, web scraping, data analysis, bot, commercial websites.

Introduction

Today, e-commerce has strong positions in the structure of world trade. Conducting trades and related financial transactions on the Internet can benefit all participants in a commercial relationship. Statistics published by a number of global financial institutions show a quantitative increase in digital buyers and sales worldwide. In the context of the current Covid-19 pandemic, the issue of the relevance of this type of trade is especially acute. The increase in the number of diseases, government measures related to restrictions on the operation of infrastructure facilities have led to a sharp increase in the fundamental indicators of e-commerce. The consequence of this surge is the emergence of the need to develop specific programs and algorithms for the convenience of shopping on the Internet.

Thus, the purpose of this article is to study the possibility of using dynamic data scraping technologies to form your own database of online stores. This project is based on testing the hypothesis about the decency of online store owners. As a rule, during the holidays, prices for some goods are reduced in order to make the buyer want to buy them. But whether the indicated price is a discount, or a fiction, we can check by developing a bot to collect data.

Theoretical research

As we stated earlier, each member of trade wins by making a purchase online. For example, buyers can significantly save time, the geography of their purchases is much wider, payment is made by card, etc. Also, buyers can become sellers themselves by placing an ad on the electronic platform (C2C). In turn, companies are paid for the

placement by providing the seller with this opportunity (B2C). In addition, the platform can make money through advertising. Thus, all transactions are transparent and tax is paid on each. In this case, the state wins.

This favorable climate leads to an increase in the influence of e-commerce in international trading. According to statistics compiled by «Statista» [1], we can look at the dynamics of global retail sales over the internet in 2014-2023. The data are represented in billion U.S. dollars.

As we can see, there has been a gradual increase in financial sales volume from 2014 to 2019. However, since 2020, the growth rate has changed significantly. This is primarily due to the current pandemic, which has made significant adjustments to all areas of economic activity.

While some companies are suffering losses that can lead to bankruptcy, others are getting stronger than before. Notable examples are Amazon, Alibaba, Prosus, whose market caps as of September 2020 are 1597.2; 777.5 and 146.1 billion dollars, respectively [2].

This year was particularly successful for Amazon, which, due to the pandemic and an increase in the share of retail sales via the Internet, earned 609 B since the beginning of the year (64.5% of the initial amount) [3]. These indicators show the potential of increasing the share of e-commerce in the global retail sales. According to experts, for the period 2021 it will be 18.1% [4].

Thus, the significant expansion of the field of e-commerce in the current and future leads to a demand for the creation of diverse programs [5]. The main task of the developers is to improve the quality of services provided and the convenience of making purchases on the Internet.

Today there is a huge variety of both separate tools for parsing data and full-fledged software. As well the use of these services is not a cheap pleasure, which is especially noticeable for medium and small businesses. Among these are "Mozenda", "Octoparse", "Data Crops" and others [6]. Therefore, one of the goals of the work is to develop the own analogue for solving similar tasks.
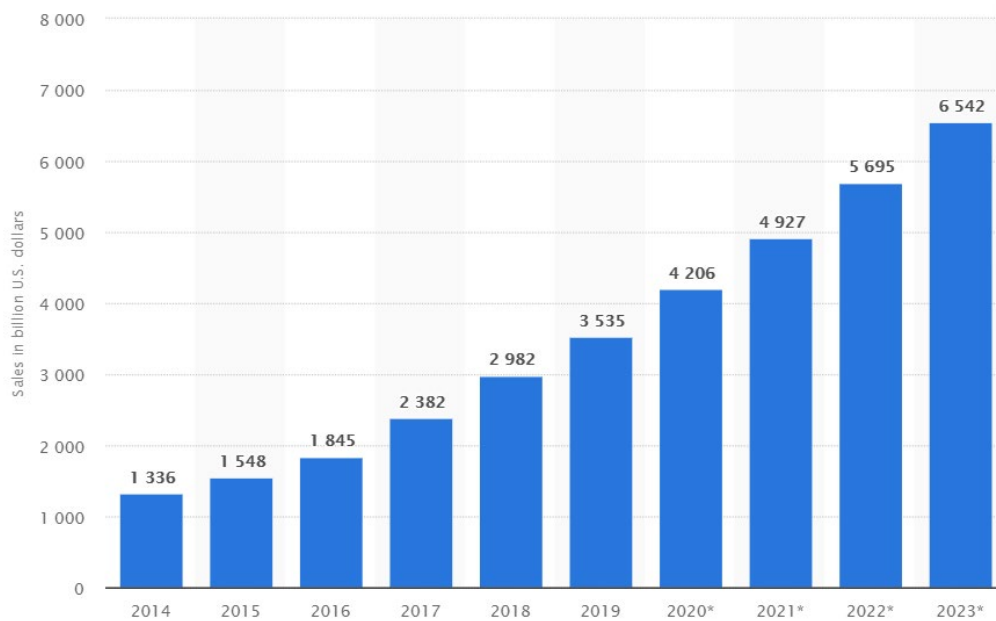
Figure 1. Retail e-commerce sales worldwide from 2014 to 2023 (in billion U.S. dollars)
Source: Developed by the author and based on [1]

Methodology

Web scraping is a technology for scanning data from web pages. By defining a number of tags and their attributes, we are able to extract the information they have. Depending on the type of information: static or the one that changes under the influence of JavaScript, web scraping is divided into static and dynamic.

In the course of this work, the Selenium WebDriver software library will be used to retrieve data from web resources. It is sometimes mistakenly called the "browser driver", but in fact it is a whole family of drivers for different browsers. It also includes a set of client libraries in various programming languages.

Often, these needs using python language, because it is more popular and has a wide scope. However, the R language has a clear advantage in the number of data science libraries, which is useful for further information processing. Therefore, writing an automated algorithm will be carried out on it.

Linking R with "Selenium WebDriver" is carried out using the "RSelenium" library. An important component for getting the job done is using Docker. This program is a tool for packaging, shipping and launching applications. Its use allows you to avoid a

number of errors that can occur when running a program or code on different computers. This property is primarily due to the structure of Docker itself and its components: "image" and "container". In order to visually see what is happening in the Docker container, the TightVNC program is used.

Table 1

Technical list of required programs and components

R language	https://cran.r-project.org/bin/windows/base/
R Studio	https://rstudio.com/products/rstudio/download/
Docker Desktop	https://docs.docker.com/docker-for-windows/install/
ChromeDriver	https://chromedriver.chromium.org/downloads
TightVNC	https://www.tightvnc.com/download.php

Source: Developed by the author

Thus, having configured all the necessary components (Table 1), we can proceed to the practical part and the analysis of the results.

Conclusions

To collect information, the electronic platform "Wine & Food" was used, where the emphasis was on the wine assortment. In the process of comparing real and discounted prices, no discrepancies were found. This fact testifies to the decency of the owners and their honest approach to buyers.

This development is quite relevant against the background of the growth of the e-commerce industry in the world. This approach is cost-efficient and allows you to track prices on online resources. It can be useful for both business and ordinary buyers. In fact, this technology can be improved to analyze a wider range of electronic platforms.

References

1. Retail e-commerce sales worldwide from 2014 to 2023 (in billion U.S. dollars). Global Business Data Platform «Statista» [An electronic resource]. – URL: https://bit.ly/3m5SMt2

2. Market cap of leading consumer internet and online service companies worldwide as of September 2020 (in billion U.S. dollars). Global Business Data Platform «Statista» [An electronic resource]. – URL: https://bit.ly/3pX1PyO

3. Amazon.com Inc (AMZN) Market Cap. Research platform «YCharts» [An electronic resource]. – URL: https://ycharts.com/companies/AMZN/market_cap

4. E-commerce share of total global retail sales from 2015 to 2023. Global Business Data Platform «Statista» [An electronic resource]. – URL: https://bit.ly/2KDFJ4o

5. Number of digital buyers worldwide from 2014 to 2021(in billions). Global Business Data Platform «Statista» [An electronic resource]. – URL: https://bit.ly/3pZu6oE

6. Top 10 Price Monitoring tool in 2020. Octoparse web scraping blog [An electronic resource]. – URL: https://www.octoparse.com/blog/top-10-price-monitoring-tool