Article

DeskDealFinder: Application for Online Shopping Optimization on Desktop Products with Web Scraping

Ivan Acero [*]

Article

DeskDealFinder: Application for Online Shopping Optimization on Desktop Products with Web Scraping

Acero Acero Ivan Ronaldo

Facultad de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano, Puno, Perú; ir.acero@est.unap.edu.pe

Abstract: Virtual commerce has experienced significant growth in recent years, especially in emerging markets such as Peru. This boom has led to an increase in the number of online stores and with it a dispersion of prices which makes it difficult for consumers to make decisions. Comparing prices has become a crucial need to maximize savings and optimize online shopping. This platform uses web scraping techniques to collect price data of desktop products from various Peruvian online stores, presenting the information in a clear and accessible way for users. By centralizing and comparing prices, it not only facilitates purchasing decisions, but also promotes transparency and competitiveness in the market. For this according to the metrics used show 56% of success of the algorithm.

Keywords: desktop products; online stores; web platforms; web scraping

1. Introduction

In Peru, e-commerce has taken a prominent place in the economy. Consumers can purchase products from the comfort of their homes and compare prices with just a few clicks. This boom has increased the number of virtual stores, leading to a price dispersion that makes it difficult for consumers to make decisions. In this context, the need arises for a tool that facilitates price comparison in an effective and efficient way.

DeskDealFinder is a platform designed to optimize online shopping for desktop products. Using web scraping techniques, it collects price data from various Peruvian online stores and presents this information in a clear and accessible way for users. By centralizing and comparing prices, the platform not only simplifies purchasing decisions, but also promotes transparency and competitiveness in the market.

Web scrapping automates the manual process of human visits to lists of web sites periodically to search and store data. Copying the data manually and pasting it into files is a tedious and time-consuming job. Automated web scrapping tool does the same job in a fraction of the time. Web scrapping software can be configured to work with any website or can be custom built for a specific website. However, generic web scrapping software may not offer any option to extract the required content due to its fixed template.[1]

The rapid growth of the Internet has led to an exponential increase in the digitization, making it difficult for users to find relevant information efficiently. Information retrieval systems are information retrieval systems are instrumental in addressing this challenge by organizing and retrieving relevant data from vast and diverse web resources. As as a technique for automatically extracting data from web pages, web scraping has become indispensable in the development of robust and efficient information retrieval systems. and efficient information retrieval systems. [2]

The main element of this is a Web Scraper (Web Scraping) that uses the company's website URL as a starting value to completely traverse the company's website domain and extract text from the pages and associated hyperlinks. The company websites are crawled using specially developed crawlers written in Python that recursively traverse the entire website. recursively traverse the entire website and collect text and link data from each web page. The crawlers are configured to collect only text data and ignore images, PDF files and other multimedia files. Since a considerable portion of website content may be stored on dynamically generated content pages when visiting a new page on the company's website. page on the company's website. [3]

Web page content stores data in a dedicated database. As a result, from the disaggregated dataset where a unit of analysis is a single URL that represents the company as a whole, we build a dataset that is as a whole, we build an overall company-level dataset. This dataset contains all the raw data and therefore raw data and, therefore, the indicators that are most relevant for the analysis. The structured schema ensures interoperability between data sources [3].

Webscraping, as described in this article, cannot be applied to applications. applications, it only applies to Internet pages. [4]

Companies and digital businesses rely on recommender systems to increase their profits, and we, as users, rely on them because they reduce information overload and make our lives easier.[5]

First, each of the data sources is accessed in turn and new links or records are identified, before downloading or deleting all available data from the identified records and links. This process is scheduled to occur daily so that new records can be captured as quickly as possible after they are posted. This reduces the risk of web links being modified or deleted before they are captured. [6]

It is important to note that this data set can be used for can be used for tasks such as classification and natural language processing. natural language processing. [7]

We use standard web scraping techniques and limit our requests in order to minimize any impact on the platform. Our goal is to collect data responsibly. [8]

the fundamental role of web scraping to drive innovation, enable more effective management of human capital dynamics and improve results. [9]

The sheer volume of online data makes manual collection and processing impractical for researchers. manual processing impractical for researchers. With the proliferation of online databases, the Internet becomes a critical resource that accentuates a critical resource that accentuates the need for fast and accurate data retrieval. The solution is called web scraping. Web scraping is the scraping is the process of implementing algorithms. [9]

2. Justification

With the growing consumption of online shopping, consumers face a challenge when trying to compare prices of similar products in different stores, which can lead to poor purchasing decisions. DeskDealFinder addresses this issue by collecting and comparing prices, offering an accessible solution for users.

In addition, the platform contributes to market transparency, allowing retailers to adjust competitively and benefiting consumers with better deals. This transparency can also incentivize stores to maintain competitive prices and improve service quality.

This project implements the web scraping technique in the Peruvian context, addressing specific challenges such as variability in web page design and legal restrictions. The integration of technologies such as HTML, CSS, PHP and Python in the development of DeskDealFinder demonstrates a practical and scalable solution that can be adapted to other contexts and markets.

Web crawling is an essential part of how the web works. It is through this process verifies whether the stores in question can be scraped as such. Therefore, companies generally do not allow web crawling. The policy of websites that cannot be scraped is usually described in a specific resource on the website called "robots.txt" file. During the web scraping process, we look for "robots.txt" files, which indicate which sections of the website are accessible to crawlers. website are accessible to crawlers. Therefore, we obey the policy described there. [3]

Acquire new data sources and methods to provide a systematic path to a fast, reliable and accurate representation of company activities, while maintaining structure and harmony. [10]

3. Methodology

The type of research for this study is applied, since it seeks to develop a practical tool to improve the efficiency of online shopping through the use of web scraping techniques. Given that it is an

applied research that should account for the level of acceptance or rejection developed by the user to the technology. [11]

The approach of this research is quantitative and descriptive. Metrics such as the number of lines of code will be used to evaluate the complexity of development and the efficiency of web scraping.

For the population and sample, research was done in virtual stores specialized in tableware.

Probably 15 virtual stores, mostly Peruvian, of Peruvian origin. The analysis of these stores will be conducted to evaluate the effectiveness of the deskdealfinder tool and price distribution.

The main objective is to evaluate the effectiveness of web scraping to capture and compare prices, and the influence of deskdealfinder on users' purchasing decisions.

The study will use a sample of 3 online stores as a guide, to ensure diversity in size and popularity.

The requirements were examined, taking into account the needs of consumers who purchase tableware online through online stores located in Peru. Speed and accuracy were determined by both functional requirements, such as the ability to extract and compare prices, and non-functional requirements.

For the instruments and techniques, Peruvian stores focused on desktop products will be the main source of data collection, with priority given to Python-based web scraping software.

Structured observation will be used. Web scraping will be used to extract products and prices and a questionnaire will be used to assess user satisfaction and perception using deskdealfinder.

Data collection will be pilot tested and a consistency check between different data sources will be performed to ensure accuracy and reliability. In addition, the effectiveness of the code will be evaluated using metrics.

The procedure for the development and evaluation of DeskDealFinder included several key steps, detailed below:

A detailed requirements analysis was carried out, considering the needs of consumers who buy desktop products online in Peruvian virtual stores. Functional requirements, such as the ability to extract and compare prices, and non-functional requirements, such as system speed and accuracy, were identified.

The MoSCoW method was used to prioritize the requirements:

- Must have: Ability to extract price data from multiple stores, friendly user interface, database to store the data.
- Should have: Filters to compare products, support for multiple product categories.
- Could have: Integration with payment platforms, personalized recommendations.

SCRUM was adopted as the agile methodology to manage the project. The main activities included:

- Sprints: 4 7-week sprints were planned to develop increments of software functionality.
- Sprint review: At the end of each sprint, deliverables were reviewed with stakeholders and feedback was obtained.
- Retrospectives: Evaluations at the end of each sprint to identify areas for improvement.

To evaluate the efficiency of code development, the following metrics were used:

- Average Processing Time: The amount of code written was measured to evaluate productivity.
- Success Rate per Page: They were counted to ensure good code documentation.
- Response Rate per Store: Monitored to evaluate code structure and readability.

The following key functionalities were developed and evaluated:

- Data Extraction: Use of *web scraping* techniques to collect pricing data of desktop products from several Peruvian online stores.
- User Interface: Development of a user friendly interface using HTML, CSS and PHP to improve the user experience.
- Database: Implementation of PhpMyAdmin to store and manage the collected data.

3.1. Análisis de datos

Data collected through web scraping will be analyzed using spreadsheets. The data will be categorized based on the order and content of relevant variables, such as product price, store and date of collection.

Once the data is obtained, a descriptive analysis will be performed to provide an overview of price dispersion and market trends.

A descriptive table of the variables used in the study is presented below:

Table 1. Variables used in the analysis

Variable	Description
Success Rate	Peruvian web sites analyzed
Success Rate per Page	Percentage of success of scraping per Page
Average Processing Time	How long the algorithm took to scrape per page
Number of Successful Attempts per Site	How many scraped products were successfully obtained

The data analysis will conclude by identifying price trends and savings opportunities for consumers. This will be based on price dispersion metrics and the efficiency of *web scraping* in gathering accurate and up-to-date data.

3.2. Ethical considerations

Since the data collected came from online stores and not from individuals, no additional ethical considerations were required. But we did take into account their privacy policy and whether they had protection which was not skipped in this research.

4. Results

Scraping Success Rate by Site

The scraping success rate per site indicates the percentage of websites where the code was able to extract data relative to the total number of sites analyzed, taking into account their policies.

Formula:

The success rate per site is calculated as:

$$\text{Success Rate by Site} =$$

$$\left(\frac{\text{Number of Sites Where Data Could Be Extracted}}{\text{Total Sites Analyzed}} \right) \times 100$$

Shops data:

- 15 Peruvian stores were analyzed considering their policies.
- 2 of them were successfully scraped.

$$\text{Success Rate per Site} = \left(\frac{2}{15} \right) \times 100 = 13.33\%$$

Given the policies of the stores and the adjustment according to the algorithm used, we were only able to apply the scraping technique to two stores.

Scraping Success Rate per page

The scraping success rate per page indicates the percentage of pages within the sites where data could be successfully extracted.

Fórmula:

The success rate per page is calculated as:

$$\text{Scraping Success Rate per Page} =$$

$$\left(\frac{\text{Number of Pages with Successful Extraction}}{\text{Total Pages Analyzed}} \right) \times 100$$

Scraped Pages Data:

- In the first store 500 pages were scraped and in the second store 1000 pages were scraped, giving a total of 1500 pages.
- 161 pages were successfully extracted from the first store and 790 pages were successfully extracted from the second store, giving a total of 951 pages successfully scraped.

$$\text{Success Rate per Page} = \left(\frac{951}{1500} \right) \times 100 = 63.4\%$$

Therefore, the scraping success rate per page in the two stores was 63.4 percent.

Average Processing Time

The average processing time indicates how long it took on average to process each web page during scraping.

Formula:

The average processing time is calculated as:

$$\text{Average Processing Time} =$$

$$\frac{\sum \text{Total Number of Pages Processed}}{\text{Total Number of Pages Processed}}$$

Processing Data:

- When scraping in the first store, the algorithm took 14 minutes and 50 seconds.
- In the second store, the algorithm took 33 minutes and 50 seconds.

$$\text{Average Processing Time} =$$

$$\frac{2919.6 \text{seconds}}{1500} = 1.94 \text{ seconds}$$

For each page, it took 1.94 seconds to be able to scrape.

Number of Successful Attempts per Site

How many successful attempts were obtained in each of the stores.

Scrape attempts data:

- At the first site, data was extracted from 500 pages and success was obtained on 161 pages.
- At the second site, an attempt was made to extract data from 1000 pages and was successful on 790 pages.

$$\text{Successful Attempts by Site:}$$

$$\text{Computactus (https://computactus.com.pe/)} :$$

$$\text{Sitio 1:} \quad \frac{161}{500} \times 100 = 32.2\%$$

Supertec (https://supertec.com.pe/) :

$$\text{Sitio 2:} \quad \frac{790}{1000} \times 100 = 79\%$$

Given the algorithm, there was more efficiency in the second store with a total of 79% and less efficiency in the first store with a total of 32%. With this, we can say that the algorithm had a 55.6% success rate scraping these pages.

5. Discussions

We wonder about the integration of gamification into current and future digital trends. [12] As the amount of information available on the web increases, so does the task of locating and analyzing it, so performing this task manually can be costly in terms of the time and effort invested. Although search engines and database engines can help to find the required information, in large digital infrastructures where search results are in the thousands - or more - new tools are needed to obtain the searched content effectively. This paper proposes the application of Web Scraping. [13]
The main web scraping tools available in the market and compares their features and functionalities. A specific tool is selected to demonstrate its use in obtaining data. The use of digital techniques such as web scraping for massive data de-loading. Thanks to this technique, unstructured information is converted.[14]
Web pages are good tools for communication and for sharing information with others, but also for participation. The method is developed in two stages (i) web scraping, which allows to obtain the links of information from web pages and to download their data, and (ii) data analysis with data cleaning. [15] In 2018, a group of researchers from the Technological University of Panama, created an algorithm using the R language to extract data from GS using the Web Scraping technique. from GS using the Web Scraping technique. [16]The algorithm, which integrates several functions, uses this unstructured data mining technique to scan the code HTML code of a web page and dynamically extracts the data the data that is displayed structured or unstructured and transforms it into table format for into table format for debugging, displaying, accessing and exporting. and subsequent export. [17] There is data that was captured by web scraping and processed using algorithms and techniques for the using algorithms and techniques for the analysis of massive network data sets. [18].
Its objective is to serve as a navigable index for the quick location of information of interest to readers, as well as a visual map that highlights the characteristics of the knowledge network that these data form through the interconnection of relationships between its nodes. This device has been created by extracting data from the magazine's website using web scraping programs. [19] In a web scraping performed from January 1 to March 31, 2015, 1 222 320 news items were captured which demonstrates the massive amount of data that can be obtained. [20] Defined queries are executed sequentially using the method of web-scraping method, until all search possibilities are exhausted. [21] First of all, it must have starting links for each platform, from which it makes a recursive search of all the courses it finds, and stores them in a list so that they can be subsequently worked by the web scraper. later worked by the web scraping. [22] This document illustrates the use of data science techniques for the creation of a database with information on prices and sales characteristics. [16]

References

1. Pillai, P.; Amin, D. Understanding the requirements Of the Indian IT industry using web scrapping. *Procedia Computer Science* **2020**, *172*, 308–313. 9th World Engineering Education Forum (WEEF 2019) Proceedings : Disruptive Engineering Education for Sustainable Development, doi:https://doi.org/10.1016/j.procs.2020.05.050.
2. Pichiyan, V.; Muthulingam, S.; G, S.; Nalajala, S.; Ch, A.; Das, M.N. Web Scraping using Natural Language Processing: Exploiting Unstructured Text for Data Extraction and Analysis. *Procedia Computer Science* **2023**, *230*, 193–202. 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023), doi:https://doi.org/10.1016/j.procs.2023.12.074.

3.   Ashouri, S.; Suominen, A.; Hajikhani, A.; Pukelis, L.; Schubert, T.; Türkeli, S.; Van Beers, C.; Cunningham, S. Indicators on firm level innovation activities from web scraped data. *Data in Brief* **2022**, *42*, 108246. doi:https://doi.org/10.1016/j.dib.2022.108246.

4.   Kempny, C.; Brzoska, P. Anwendungskontexte von Web Scraping in der Versorgungsforschung - Nur für Web-Expert:innen? Oder eine Methode für alle Versorgungsforscher:innen!? *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* **2023**, *176*, 61–64. doi:https://doi.org/10.1016/j.zefq.2022.11.010.

5.   Sakieh, Y. Shaping climate change discourse: the nexus between political media landscape and recommendation systems in social networks. *Social Network Analysis and Mining* **2024**, *14*. Cited by: 0; All Open Access, Hybrid Gold Open Access, doi:10.1007/s13278-023-01175-9.

6.   Wyatt, F.; Robbins, J.; Eaton, S. Implementing a routine and standard approach for the automatic collection of socio-economic impact observations for impact-based forecasting and warning. *International Journal of Disaster Risk Reduction* **2024**, *110*, 104608. doi:https://doi.org/10.1016/j.ijdrr.2024.104608.

7.   dos Reis Filho, I.J.; de Campos Coleti, J.; Marcacini, R.M.; Rezende, S.O. Dataset: Annotated soybean market news articles. *Data in Brief* **2024**, *55*, 110545. doi:https://doi.org/10.1016/j.dib.2024.110545.

8.   Yasin, A.; Fatima, R.; Ghazi, A.N.; Wei, Z. Python data odyssey: Mining user feedback from google play store. *Data in Brief* **2024**, *54*, 110499. doi:https://doi.org/10.1016/j.dib.2024.110499.

9.   Goulas, S.; Karamitros, G. How to harness the power of web scraping for medical and surgical research: An application in estimating international collaboration. *World Journal of Surgery* **2024**, *48*, 1297 – 1300. Cited by: 0; All Open Access, Hybrid Gold Open Access, doi:10.1002/wjs.12220.

10. Hajikhani, A.; Pukelis, L.; Suominen, A.; Ashouri, S.; Schubert, T.; Notten, A.; Cunningham, S.W. Connecting firm's web scraped textual content to body of science: Utilizing microsoft academic graph hierarchical topic modeling. *MethodsX* **2022**, *9*, 101650. doi:https://doi.org/10.1016/j.mex.2022.101650.

11. Muñoz Bonilla, H.A.; Vasco Gutiérrez, D.F. Contributions for the evaluation of gamified pedagogical strategies with serious games and intervention of luck; [Aportes para la evaluación de estrategias pedagógicas gamificadas con juegos serios e intervención del azar]. *Revista Interuniversitaria de Formacion del Profesorado* **2024**, *99*, 231 – 252. Cited by: 0; All Open Access, Gold Open Access, doi:10.47553/rifop.v99i38.1.93896.

12. Moreno, C.B.; Carretero, M.R.M.; de Santiago, B.S.R.; Rumayor, L.R. Gamification-Education: the power of data. Teachers in social networks; [Gamificación-educación: el poder del dato. El profesorado en las redes sociales]. *RIED-Revista Iberoamericana de Educacion a Distancia* **2024**, *27*, 373 – 396. Cited by: 0; All Open Access, Gold Open Access, doi:10.5944/ried.27.1.37648.

13. Aguilar, H. Scraping Archaeology: A Methodological Approach from the Web Scraping and Text Mining; [Raspando la Arqueología: Una Aproximación Metodológica desde el Web Scraping y Text Mining]. *Revista del Museo de Antropologia* **2023**, *16*, 439 – 450. Cited by: 0; All Open Access, Gold Open Access, doi:10.31048/1852.4826.v16.n2.41094.

14. Escandell-Poveda, R.; Papí-Gálvez, N.; Iglesias-García, M. Digital techniques for the study of professional skills and profiles: the case of SEO job offers; [Técnicas digitales para el estudio de las competencias y perfiles profesionales: el caso de la oferta laboral de SEO]. *Scire* **2023**, *29*, 31 – 42. Cited by: 2; All Open Access, Hybrid Gold Open Access, doi:10.54886/scire.v29i1.4877.

15. Rosso-Mateus, A.E.; Montilla-Montilla, Y.M.; Garzón-Martínez, S.C. Methodology for the Collection and Analysis of Real Estate Data Using Alternative Sources: Case Study in Three Medium-Sized Cities of Colombia; [Metodología para obtención y análisis de datos inmobiliarios usando fuentes alternativas: estudio de caso en tres ciudades intermedias de Colombia]. *Ingenieria (Colombia)* **2022**, *27*. Cited by: 0; All Open Access, Gold Open Access, doi:10.14483/23448393.17952.

16. Rubio, J.A.C.; Guzmán, F.J.C.; Otero, J. An internet-based data set of prices and characteristics of dwelling in Colombia; [Construção de uma base de dados de preços e características de moradia para a Colômbia]; [Una base de datos de precios y características de vivienda en Colombia con información de internet]. *Revista de Economia del Rosario* **2019**, *22*, 75 – 100. Cited by: 0; All Open Access, Gold Open Access, Green Open Access, doi:10.12804/revistas.urosario.edu.co/economia/a.7768.

17. Murillo, D.; Saavedra, D.; Zapata, R. Web application in Shiny for the extraction of data from profiles in Google Scholar; [Aplicación web en Shiny para la extracción de datos de perfiles en Google Scholar]. *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology* **2022**, *2022-July*. Cited by: 1; All Open Access, Bronze Open Access, doi:10.18687/LACCEI2022.1.1.235.

18.   Zarrabeitia-Bilbao, E.; Morales-I-gras, J.; Rio-Belver, R.M.; Garechana-Anacabe, G. Green energy: Identifying development trends in society using Twitter data mining to make strategic decisions; [Energía verde: Identificación de tendencias en la sociedad mediante la minería de datos aplicada a Twitter para la toma de decisions estratégicas]. *Profesional de la Informacion* **2022**, *31*. Cited by: 8; All Open Access, Bronze Open Access, doi:10.3145/epi.2022.ene.14.

19.   Gonzales, A.; Colmenero-Ruiz, M.J.; Pinto, A.L. A cartography of the Profesional de la información journal: a visual map of 30 years of history; [Cartografía de la revista Profesional de la información: mapa visual de 30 años de historia]. *Profesional de la Informacion* **2021**, *30*. Cited by: 0; All Open Access, Bronze Open Access, doi:10.3145/epi.2021.nov.05.

20.   Cobos, T.L. Journalism industries in the internet era: The case of Colombian news media outlets in Google News Colombia; [Indústrias jornalísticas na era da internet: O caso da mídia Colombiana no Google News Colombia]; [Las industrias periodísticas en la era de internet: El caso de los medios noticiosos colombianos en Google News Colombia]. *Contratexto* **2020**, p. 85 – 104. Cited by: 1; All Open Access, Gold Open Access, Green Open Access, doi:10.26439/contratexto2020.n033.4785.

21.   Blázquez-Ochando, M.; Ramos-Simón, L.F. Digitization of protected works: Software for the detection of out of commerce works; [Digitalización de obras protegidas: Software para la detección de obras fuera del circuito comercial]. *Profesional de la Informacion* **2019**, *28*. Cited by: 1; All Open Access, Bronze Open Access, doi:10.3145/epi.2019.nov.24.

22.   Jerson Erick Herrera Rivera, B. Recommender system using web scraping for enrollment in MOOCs of students in engineering careers at the Public University of Arequipa; [Sistema de recomendación usando web scraping para matrícula en moocs de estudiantes en carrera de ingeniería en universidad pública de arequipa]. *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology* **2019**, *2019-July*. Cited by: 0; All Open Access, Bronze Open Access, doi:10.18687/LACCEI2019.1.1.42.