IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

Cloud Based Web Scraping Automation

Lavanya [1], Leelavathi N[2], Manoj M[3], Nancy K[4], Mohammad Khurram J[5]

Student, Department of Computer Science and Engineering, Dayananda Sagar University, Bengaluru, Karnataka, India[1234]

Professor, Department of Computer Science and Engineering, Dayananda Sagar University, Bengaluru, Karnataka, India[5]

ABSTRACT: The main objective of this paper is to provide upgraded web scraping services by addressing crucial gaps in the current market space. Our platform offers personalized news aggregation with multi language support and summary tools, catering to personalized preferences. Unlike existing summary news providers that offer generalized content, this platform allows users to customize news updates and summaries on very specific topics. It also includes features such as metadata scraping and processing of relevant YouTube videos. The system thus provides a level of customization unprecedent in the market. Features such as smart summary and multi-language translation are embedded within the services. Moreover, the project provides tools for data analysts with statistics and graphical representations derived from scraped product information. Adding to this, the information access Application is platform independent and can be used on Android, Windows, MacOS, Linux systems. The user can disable any of the services at any given moment using the desktop application and can also adjust the frequency at which the scraping is performed. This project also intends to automate every day, tedious tasks such as news gathering based on keywords or product price checking at regular time intervals by combining various technologies of Computer Science Engineering namely Web Scraping, Cloud Data access and Retrieval, Cloud Server Hosting, Real time Database Service, Desktop Application Development, Tool Automation.

KEYWORDS: Web Scraping, Automation, Real Time Database.

I. INTRODUCTION

The internet serves as an abundant source of valuable information across diverse domains. However, accessing and extracting data from websites manually can be time-consuming and inefficient. Web scraping, a technique used to automatically gather data from web pages, has emerged as a powerful solution to this challenge.This abstract presents an overview of web scraping, including its methodologies, applications, challenges, and ethical considerations. The methodology section discusses various techniques employed in web scraping, such as parsing HTML/XML documents, utilizing APIs, and employing browser automation tools like Selenium. Moreover, it explores the importance of choosing appropriate scraping libraries and frameworks based on project requirements and target websites. The applications section highlights the versatility of web scraping across different industries and domains. It outlines its utility in market research, competitive analysis, lead generation, sentiment analysis, and more. Additionally, it discusses its role in academic research, including data collection for social science studies and sentiment analysis of online discourse. Challenges associated with web scraping, such as dynamic website content, anti-scraping measures, and legal concerns, are addressed in the subsequent section. Strategies to mitigate these challenges, such as implementing delays, rotating proxies, and respecting website terms of service, are also discussed.Lastly, ethical considerations surrounding web scraping are examined, emphasizing the importance of respecting website owners' rights and privacy concerns. It underscores the need for transparency, accountability, and adherence to ethical guidelines while scraping data from online sources.

In conclusion, this abstract provides a comprehensive overview of web scraping, elucidating its methodologies, applications, challenges, and ethical considerations. It serves as a valuable resource for researchers, practitioners, and enthusiasts seeking to harness the power of web scraping for data extraction purposes.

II. RELATED WORKS

[1]. This Chapter details about with the advent of new technologies, internet usage has skyrocketed and the amount of unstructured information that internet users generate online has increased. Web scraping plays an important role for extracting this unstructured information from the internet. Their approach aims to solve the problem of scraping as well

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

as the viability of large data application for data-based industries. They used Selenium which offers web drivers to stimulate a real user interface with browser, and AWS like Elastic Compute Cloud or DynamoDB for compute and storage for the data generated.

It is widely acknowledged that the internet and e-commerce have predominantly thrived in developed nations. Daithiya Sudan, Perumalraja and Kamalesh [2] clearly discussed that careful handling of text formatting during data scraping is important to retrieve accurate results, while addressing trustworthiness concerns through Amazon's upvote system. Cleaning procedures involves removing special characters for extracting data for extracting user review, helpful count. After classification of reviews into service, feature or product categories, the processed output need to be consolidated into a single CSV file and load into database for visualization and summarization purpose.

[3] In today's world the need for online shopping has increasing day by day. Web scraping is generally a new strategy for gathering the web information especially, product analysis which helps the online users to grab the best deals for their product from multiple-ecommerce websites. Using a python and BeautifulSoup is a straightforward way to collect data from multiple websites had tailor accordingly to choose websites available for desired product which is used for data analysis. YouTube, the world's largest video-sharing platform, is the goldmine of information and data.

[4] This paper published about scraping of data from the youtube, the data may be video titles, description, tags, view counts, likes, comments, and other metadata. They have used tools like Octoparse, import.io, Grepsr. This tool helps to extract data and save It into various formats like csv, excel or Json.

[5] Wu-Jeng LI, Chiaming Yen, You-Sheng Lin, Shu-Chu Tung provided JustIoT Internet of Things based on the Firebase Real-time Database proposed that JustIoT's architecture offers significant utility for web scraping purposes, by leveraging backend Google Firebase.Real-time database, as it provides a intelligent sever equipped with MQTT connection support and condition control capabilities, which in turn enhances the efficiency and accuracy of web scraping tasks. The integration of single-page applications facilitates smooth data retrieval and manipulation, which empowers users to extract valuable information from diverse online sources with ease. The proliferation of news content online has led to challenges in accessing articles form websites.

[6] The author presented to scrape data form three news websites: Detik, Tribunnews and Liputan6. Their approach involves devising Regrex patterns and implementing this as rules to scrape the content. And, to filter the patterns to eliminate non-news elements like advertisements and scripts.
In the overview of the research and practices, web scraping is a powerful tool for extracting and analysing data from vast expanse of internet. Web scraping has become an essential tool for businesses, research and individual seeking to extract valuable insights and data from the platform. As the digital landscape continues to evolve, understanding and harnessing the capabilities of web scraping will remain essential for unlocking actionable insights and gaining a competitive edge in an increasingly data-driven world.

III. DESIGN

Cloud web scraping involves setting up cloud infrastructure, developing scraping scripts, and storing scraped data in cloud storage. The process includes distributed processing to handle largescale scraping tasks efficiently and data processing for cleaning and normalization. Monitoring and compliance ensure the scraping process operates smoothly and ethically, while regular maintenance optimizes performance website changes.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

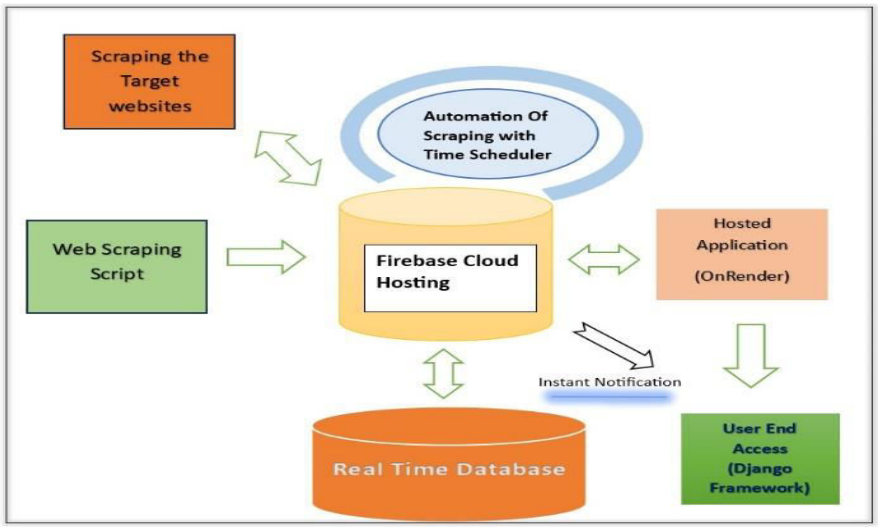


Fig 1: Block diagram of proposed model

In the above figure, Cloud-based web scraping automation harnesses the power of cloud computing to efficiently extract data from websites and online sources. Leveraging Python libraries like BeautifulSoup (BS4), Selenium, and Newspaper3, it enables seamless scraping of web content across multiple platforms. Utilizing OnRender for deployment, this solution offers free hosting, ensuring accessibility and scalability.

With Python or Firebase Cloud Functions as scheduler scripts, tasks can be automated and scheduled for optimal performance. The app structure, organized into individual Django apps, facilitates modular services for enhanced flexibility and maintainability. Data access and storage are streamlined through OnRender and Firebase Realtime Database, providing secure and reliable storage solutions. Django serves as the backend framework, offering robust web development capabilities.

Built-in Django email service enables timely notifications and updates. Dependency management, handled through virtual environments and shell scripts, ensures smooth integration and operation. For large-scale scraping tasks, the YouTube Data API is employed for efficient comments retrieval. Additionally, product analysis is enhanced with the Amazon-buddy Python wrapper class, while news and summarization benefit from access to over 80,000 channels through the News API. This comprehensive cloud-based solution empowers users with efficient data extraction, analysis, and summarization capabilities, facilitating informed decision-making and insights generation.

IV. EXPERIMENTAL WORK

4.1 Youtube Analytics

4.1.1 Youtube videos Scraping

YouTube scraping is designed to collect, process, and analyze data from YouTube videos, channels, and comments. It leverages APIs provided by YouTube along with web scraping techniques to gather information efficiently and ethically. This project integrates the YouTube Data API, ensuring compliance with YouTube's terms of service, to fetch comprehensive metadata including video details, channel information, comments, and related data.

It also extracts supplementary details such as video tags and engagement metrics. The application enables users to search for YouTube videos based on keywords, utilizing the YouTube Data API to gather relevant metadata and captions. Captions are extracted and stored locally using 'captions.py', facilitating easy access and organization based on search keywords. The Sumy library is employed for text summarization, employing LSA for concise content extraction, with summaries saved as text files for each video. Furthermore, PDF summaries are generated using Jinja2 templating and WeasyPrint library,
incorporating thumbnail images, and stored in designated directories.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

4.1.2 Sentimental Analysis

our analytical capabilities extend to YouTube comment analysis, leveraging state-of-the-art algorithms such as TextBlob and VADER (Valence Aware Dictionary and sentiment Reasoner) for sentiment analysis. This methodology combines automated data retrieval with advanced analytical techniques to empower users with actionable insights derived from a wide array of online sources.

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule based sentiment analysis tool specifically developed for evaluating sentiment in text data, particularly in social media contexts. This algorithm operates by leveraging a predefined lexicon containing words annotated with sentiment scores, reflecting their positivity or negativity.

VADER computes a compound sentiment score for a given piece of text by aggregating individual word sentiment scores while considering their contextual relevance and syntactic structure. This score ranges from -1 (indicating extreme negativity) to +1 (indicating extreme positivity), with values near zero suggesting neutrality. Latent Semantic Analysis (LSA) is a method used in natural language processing (NLP) to extract underlying semantic structures from textual data. It operates by representing documents and terms as vectors in a high-dimensional space, where each dimension corresponds to a unique term in the vocabulary.

4.2 Product Analytics

Web scraping for product analytics involves gathering data from various online sources to analyze product performance and trends. By taking inputs such as minimum price, maximum rating, and desired number of results from the user, the system can filter and extract relevant product data from e-commerce platforms or review websites. This data could include product specifications, customer reviews, ratings, and pricing information. Additionally, the system can utilize trend analysis techniques to visualize product performance over time using graphs or charts. These graphs may depict trends in sales, customer satisfaction, pricing fluctuations, or popularity, providing valuable insights for product analysis and decision-making.

Following the comprehensive product analysis based on user requirements, the findings are distilled into a structured format for dissemination to the user. One effective method is to compile the results into a CSV (Comma-Separated Values) file, offering a convenient and universally compatible format for data representation. The CSV file can encapsulate key insights, including product performance metrics, consumer sentiment analysis results, market trends, and any other relevant findings derived from the analysis process. This ensures seamless delivery of the analyzed data to the end-user, facilitating further review, collaboration, and decision-making.

Market segmentation strategies can be employed to categorize consumers into distinct groups based on various factors such as demographics, psychographics, or behavioral patterns. This allows for a more targeted analysis of consumer behavior and preferences, enabling businesses to tailor their product offerings and marketing strategies to specific audience segments.

4.2.1 Trending Analysis

- Price Analysis: Utilizes statistical analysis and data visualization to identify pricing trends and calculate metrics such as average, minimum, and maximum prices for each product.
- Competitor Analysis: Identifies key competitors on Amazon, analyzes their pricing strategies, discounts, and promotions to assess competitiveness.
- Pricing Optimization: Recommends pricing strategies based on market demand, competitor prices, and historical sales data.
- Price Retrieval and Presentation: Utilizes the AmazonBuddy library to retrieve product prices based on user-defined search parameters, including keywords, sorting preferences, and price range.
- Data Visualization: Dynamically generates a price comparison graph using Matplotlib, plotting prices against ASINs to provide visual insights into product pricing.
- Report Generation and Export: Offers PDF reports summarizing search parameters and price comparison graphs, generated using the WeasyPrint library. Users can also export product data in CSV format for further analysis.
- User Interface for Product Search: Allows users to input search criteria through a form, retrieve product data, and view price comparison graphs directly on the web interface.
- Price Comparison Graph: Dynamically generated using Matplotlib, plotting product prices against ASINs to visually represent pricing trends and variations.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

- PDF Report Generation: Generates PDF reports containing search criteria and price comparison graphs, rendered from HTML templates and served as downloadable attachments.
- CSV Data Export: Enables users to export product data (ASINs and prices) to CSV files, generated based on search form data and served as downloadable attachments.

4.2.2 Price Analysis

- Conducts statistical analysis and data visualization to identify pricing trends and patterns.
- Calculates metrics such as average price, minimum price, maximum price, and price variance for each product.
- Compares prices across different sellers, brands, and product categories to assess competitiveness.
- Identifies outliers and anomalies in pricing data to uncover potential errors or irregularities.
- Utilizes historical pricing data to forecast future price trends and inform pricing strategies.
- Presents price analysis findings through graphical visualizations such as histograms, box plots, and scatter plots.
- Provides insights into price elasticity and demand elasticity to optimize pricing decisions.
- Incorporates machine learning algorithms to analyze large datasets and identify hidden patterns in pricing data.
- Enables users to customize price analysis parameters such as time range, product category, and geographical region.
- Facilitates real-time monitoring of pricing fluctuations and market dynamics to support agile pricing adjustments.
-

V. RESULTS

5.1 Youtube videos Scraping

Using user-defined keywords, our system efficiently scrapes YouTube videos, extracts their subtitles, and summarizes the content into a concise PDF format. This PDF summary, tailored to the user's interests, can be translated into multiple languages and is promptly emailed to the user for convenient access.

Wion latest news Summary

hurricanes floods storms drought the world woke up to freak weather almost every month in the year 2021. a report has now revealed that 10 of the most extreme weather events cost nearly 170 billion dollars in damages almost 20 billion dollars more than the year 2020. now this is also the sixth year in a row that natural disasters cost more than a hundred billion dollars worth of damage the upward trend signals add the growing impact and pace of climate change and related disasters the most expensive disaster in 2021 was hurricane ida the category 4 hurricane lashed the eastern united states also causing massive flooding in the new york area especially the hurricane claimed 95 lives in the u.s causing more than 65 billion dollars in damage the second one was the flooding situation in europe in july heavy rainfall led to flash floods across europe parts of germany france the netherlands and belgium were inundated the floods caused more than 43 billion dollars in losses the next one on the list was the winter storm in texas the storm took out the state's entire power grid leading to the largest blackouts in the u.s in 20 years the total damage has been estimated at around 23 billion dollars the fourth most expensive disaster were the floods in china's henan province more than 800 000 people were evacuated from their homes with more than 14 million people affected by these floods the total damage here was around 17.6 billion dollars the other floods causing billions of dollars in damage were the floods in canada the british columbia flooding caused an estimated damage of 7.5 billion dollars other disasters that wreak havoc across the world include a late spring freeze in france also a cyclone in india and bangladesh in may and the list technologies are biased towards richer nations especially

This streamlined process ensures that users receive relevant and digestible summaries of YouTube video content, empowering them to stay informed and engaged with their preferred topics effortlessly.

5.1.1 Youtube Sentiment Analysis

Advanced version of analysis employs two models, **TextBlob and VADER,** for emotion and sentiment analysis, respectively.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

1.1 Using VADER Algorithm

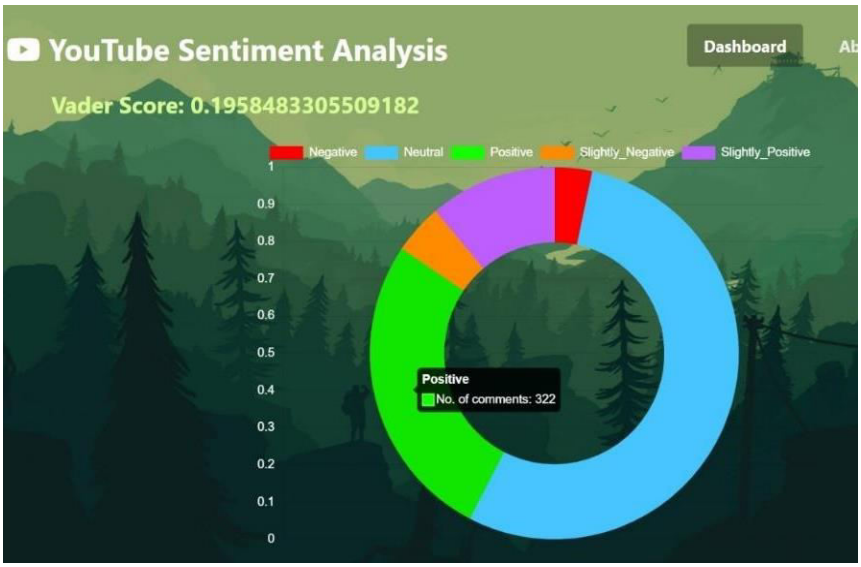


Fig 3 : Sentiment Analysis using VADER

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule based sentiment analysis tool specifically developed for evaluating sentiment in text data, particularly in social media contexts. This algorithm operates by leveraging a predefined lexicon containing words annotated with sentiment scores, reflecting their positivity or negativity. Additionally, it incorporates rules and heuristics to account for various linguistic nuances, including intensifiers, modifiers, negations, and emotive language. VADER computes a compound sentiment score for a given piece of text by aggregating individual word sentiment scores while considering their contextual relevance and syntactic structure.

1.2 Using TextBlob Algorithm

Fig 4: Sentimental analysis using TextBlob

The graph visualizes sentiment analysis results obtained using the TextBlob algorithm, which assesses the polarity of text data to determine sentiment. Each data point represents a piece of text, such as a comment or review, with its sentiment polarity plotted on the y-axis. The x-axis may represent time, document ID, or another relevant variable. Positive sentiment is typically depicted as values above the baseline, negative sentiment as values below the baseline, and neutral sentiment as points near the baseline. By analyzing the distribution and trends of sentiment polarity over

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

time or across documents, users can gain insights into audience sentiment, identify sentiment outliers, and track changes in sentiment perception.

5.1 Comments Statistics

Fig 5 Average comments per video

The above graph is calculating statistical data such as average comments per day, likes, sorting, and filtering of comments can beextracted and analyzed. This process involves collecting the total number of comments over a period and calculating the average comments per day to gauge engagement. Additionally, likes and other metrics can be gathered to assess the video's popularity and audience reception. Sorting and filtering functionalities allow for organizing comments based on relevance, sentiment, or other criteria to derive insights and enhance user experience.

5.3 Product Analysis

5.3.1 Trending products visualization

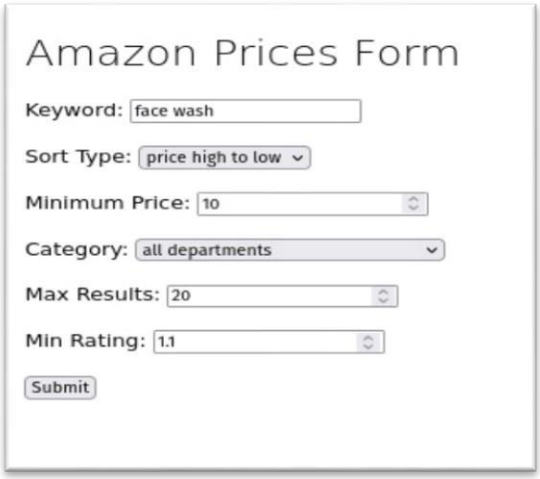
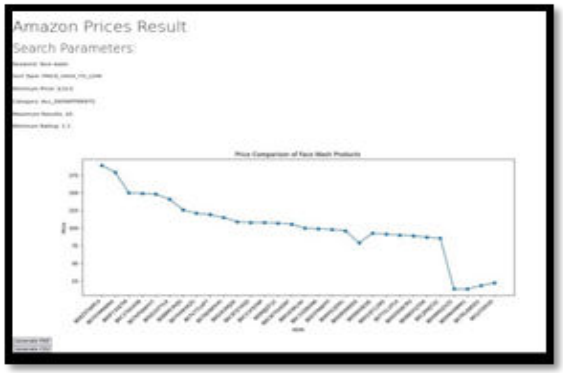


Fig 6: Trending Products Visualization

when you input a specific department and define parameters like search filters and the maximum number of results per letter, it retrieves all the trending products within that category. Once the results are obtained, it conducts a graphical analysis, mapping the search filter against the count of trending words. This visual representation provides insights into the distribution and popularity of various products within the chosen department. Moreover, users have the option to download the generated PDF and CSV files containing the detailed information and analysis for further examination and processing.

5.3.2 Price analysis based on user input



International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

Fig 7: User Input form For Amazon Prices Estimation.

The user can enter a keyword to search for a product and specify sorting preferences, choosing between sorting the product prices from high to low. Additionally, the user can set the minimum price at which the product search should begin and define the maximum number of results desired. Moreover, users
can input the minimum rating they require for the products displayed.

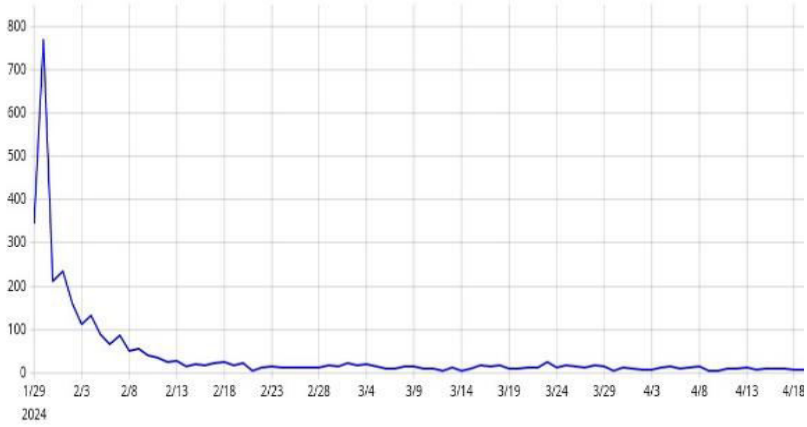


Fig 8 : ASIN vs. Price Analysis for Face Wash Products.

In the above graph users input keywords (products), select a sorting type (high to low or low to high), set a minimum price per category, and define filters like maximum results and minimum rating. The service then provides a graphical analysis depicting all the products' IDs against their respective prices. For instance, consider the graph where the x-axis represents the ASINs (Amazon Standard Identification Numbers) of different face washes, while the y-axis displays their corresponding prices. This visual representation offers users insights into the price distribution across various face wash products.

VI. CONCLUSION

Cloud-based web scraping automation offers a myriad of advantages for businesses and individuals seeking efficient and scalable data extraction solutions. By leveraging the power of cloud computing, organizations can achieve increased flexibility, scalability, and costeffectiveness in their web scraping initiatives. The ability to distribute workloads across multiple servers, access on-demand computing resources, and benefit from enhanced storage

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304450 |

capabilities contributes to improved performance and faster data retrieval. Moreover, cloud-based solutions reduce the need for extensive infrastructure investment and maintenance, allowing users to focus on developing and refining their scraping algorithms rather than managing hardware resources. The scalability of cloud platforms ensures that web scraping operations can easily adapt to varying workloads and data volumes, making it an ideal choice for projects with dynamic requirements. However, it is crucial to consider ethical and legal aspects of web scraping, respecting website terms of service and applicable laws. Transparency and responsible data usage practices are key to maintaining a positive reputation and avoiding potential legal issues. In summary, the adoption of cloud-based web scraping automation provides a strategic advantage for organizations by enhancing efficiency, scalability, and cost-effectiveness while enabling them to navigate the complexities of web data extraction in a rapidly evolving digital landscape.

REFERENCES

[1] Chaulagain, Ram Sharan, et al. "Cloud based web scraping for big data applications." 2017 IEEE International Conference on Smart Cloud (SmartCloud). IEEE, 2017.

[2] P.S.Gaikwad, Kaushal Parmar, Rohit Yadav, Datta Supekar Associate Professor ,Student, Student, Student Computer Engineering Department,IMPLEMENTATION OF WEB SCRAPING FOR E-COMMERCE WEBSITE Prof. All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune, India.

[3] Thomas, David Mathew, and Sandeep Mathur. "Data analysis by web scraping using python." 2019 3rd International conference on Electronics Communication and Aerospace Technology (ICECA). IEEE, 2019.

[4] Scott RE. Data Scraping YouTube for the Study of Lieder Reception. Nineteenth Century Music Review. 2022 Dec;19(3):655-67.

[5] Chiaming Yen, You-Sheng Lin, Shu-Chu Tung JustIoT Internet of Things based on the Firebase Real-time Database by Wu-Jeng Li, in 2018.

[6] Achmad Maududie; Windi Eka Yulia Retnani; Muhamat Abdul Rohim An Approach of Web Scraping on News Website based on Regular Expression.

[7] Khder MA. Web scraping or web crawling: State of art, techniques, approaches and application. International Journal of Advances in Soft Computing & Its Applications. 2021 Nov 1;13(3).

[8] Khatter, Harsh, Akshat Sharma, and Ajay Kumar Kushwaha. "Web Scraping based Product Comparison Model for E-Commerce Websites." 2022 IEEE International Conference on Data Science and Information System (ICDSIS). IEEE, 2022.

[9] Landu, Tony Tona, et al. "Machine learning algorithm for text categorization of news articles from Senegalese online news websites." 2022 17th Iberian Conference on Information Systems and Technologies (CISTI). IEEE 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details