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 12, Issue 5, May 2023

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 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205063 |

AI Search Engine for Clutter Free Web Research

Jagtap Gauravi Sopan, Pawar Snehal Sunil, Kanade Omkar Rajesh, Prof. K .S . Khamkar

Department of Computer Engineering, Shri Chhatrapati Shivajiraje College of Engineering, Dhangawadi, Pune, India

ABSTRACT: A search engine is a software program that helps people find the information they are looking for online using keywords or phrases. Search engines are able to return results quicklyeven with millions of websites onlineby scanning the Internet continuously and indexing every page they find. Now a day's websites are flooded with unnecessary promotional messages, ads, entertaining content it's become difficult to focus on topic while researching on web. To find meaningful information you have to visit multiple websites is time consuming. In this project, we want to create system which visit multiple websites simultaneously and understands html structure of all websites and extract only main content and later it will sort content based on their type like definitions, bullet points, table, paragraphs and important notes etc.

KEYWORDS: -Search Engine, Web Scarping, Summarization, Clutter Free, Android App, Artificial Intelligence, Ai Module

I. INTRODUCTION

Searching is one of the most used actions on the Internet. Search engines as an instrument of searching, are very popular and frequently used sites. This is the reason why webmasters and every ordinary user on the Internet, must have good knowledge about search engines and searching.Webmasters use major search engines for submitting their sites on it, and for searching. Ordinary users use major search engines primarily for searching, and sometimes for submitting their homepages or small sites. Why search engines are so popular? If you are ordinary user and you want to find some texts or pictures about certain theme, first thing you will do is to visit search engine to get some good URLs about certain theme. This will happen every time you need some information or data about any theme. If you are webmaster, you will also need some information while preparing site for the Web, and you will also use search engines. Then, when you finish with it, you must submit your URL to many search engines. After that you will check your URL ranking on every search engine. There is also hot news on every major search engine, many other interesting contents... All of this shortly describes why search engines are so popular. As a user at novice level, you must learn how to use search engines for searching the Internet. You must know that there are two ways of searching: by using user's query or by using categories. If you have keywords or phrase that best describes the theme you need, you should use user's query. But if you need some theme, and you don't have keywords or phrase, you should use categories.

II. LITERATURE SURVEY

This is also well known as Web Scraping Using Summarization. In this study, previous research related support is needed. Content Quality research is needed to get indicators as assessment material, then the Content Audit method as a stage of evaluation and validation, spell checking method as a reference automation check, and web scraping method as data collection automation.

This Paper is published 2021 covers a number of topics including web scraping, website auditing, and content quality. It emphasises the significance of website content quality and the need to take into account a variety of factors while evaluating it. The three stages of the website auditing process include data collection, analysis, and final reporting. During the data collecting stage, information is gathered from the business, the site's developer, the users, and the auditors. In the analysis step, the data that have been gathered are evaluated, and the accuracy of the information on the website is evaluated. The final report offers suggestions for raising the calibre of the content. The paper also discusses utilising a spellchecker library to find spelling mistakes and using web scraping to retrieve data from websites..

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

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205063 |

Proposed System Architecture: -

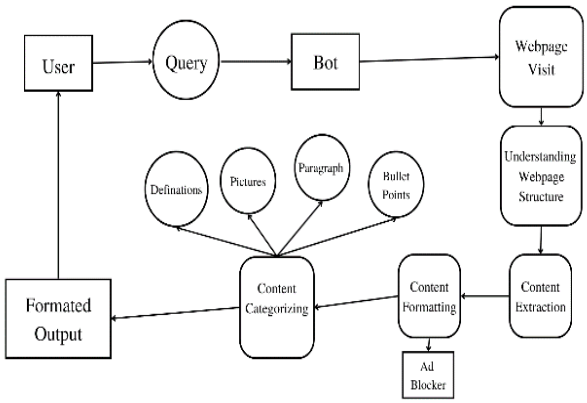

Fig. 1 Proposed System Architecture

Implementation: -

The project is implemented by using through following modules

1. Real time Search results extraction Module

An essential part of this project is the Real-Time Search Results Extraction module. This module is in charge of extracting all the URLs from the websites that show up as search results once the user types a query to conduct a search for information. The search results page is automatically browsed, and the URLs of pertinent websites are gathered. The module makes it possible for the project's next stages to analyse and evaluate each website's content by extracting the URLs. This guarantees that the app can give the user accurate and current information. By effortlessly extracting the required links for additional processing and analysis, the Real-Time Search Results Extraction module improves the usefulness of the Android app. The Real-time Search Results Extraction bot operates seamlessly in the background, swiftly and accurately retrieving URLs without disrupting the user's search experience. Its efficiency and accuracy contribute to a more efficient and productive research process, allowing users to focus on evaluating the content of the websites rather than spending time on manual URL extraction.

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

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205063 |

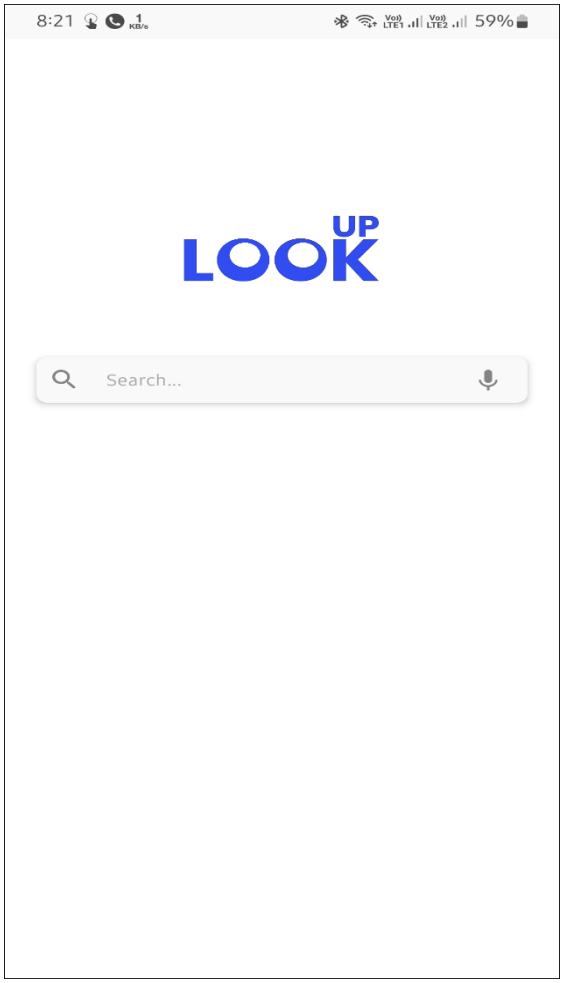

Fig. 2. Real time Search results extraction Module

2. URL Visitor Module

The URL Visitor module plays a critical role in this project by actively exploring the URLs obtained through the Real-Time Search Results Extraction module. It emulates the behaviour of a genuine user, manually visiting each URL individually. This simulated user interaction allows the module to access the webpages and retrieve the HTML content of each site. By visiting the URLs and extracting the HTML content, the URL Visitor module enables subsequent project stages to analyse and process the webpage data effectively. This includes tasks like HTML parsing, relevant information extraction, and structured storage for future use. The URL Visitor module's manual visitation of the URLs ensures the app can gather accurate and comprehensive data from each website, providing users with reliable and pertinent information. Its role is vital in the overall functionality of the project, facilitating the extraction and analysis of webpage content to create a seamless user experience.

3. Content Extraction Module

TheContent Extractor module, a vital component in this project, plays a crucial role in processing the raw HTML data retrieved by the URL Visitor module. With its advanced parsing capabilities, this module extracts key elements from the HTML, including the website's title and a concise summary of its content. By accurately identifying and isolating these important details, the Content Extractor module prepares the extracted information for seamless integration into the Database module. The resulting query, which combines the Website Title, Website URL, and Summary of the Website, ensures that the data is efficiently stored and organized in the database for easy retrieval and presentation to users. This intelligent automation significantly enhances the user experience by providing relevant and concise

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

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205063 |

information from the webpages while eliminating unnecessary clutter. The Content Extractor module's ability to process and structure the extracted content adds value to the overall functionality of the project, facilitating efficient data management and enhancing the app's usability.

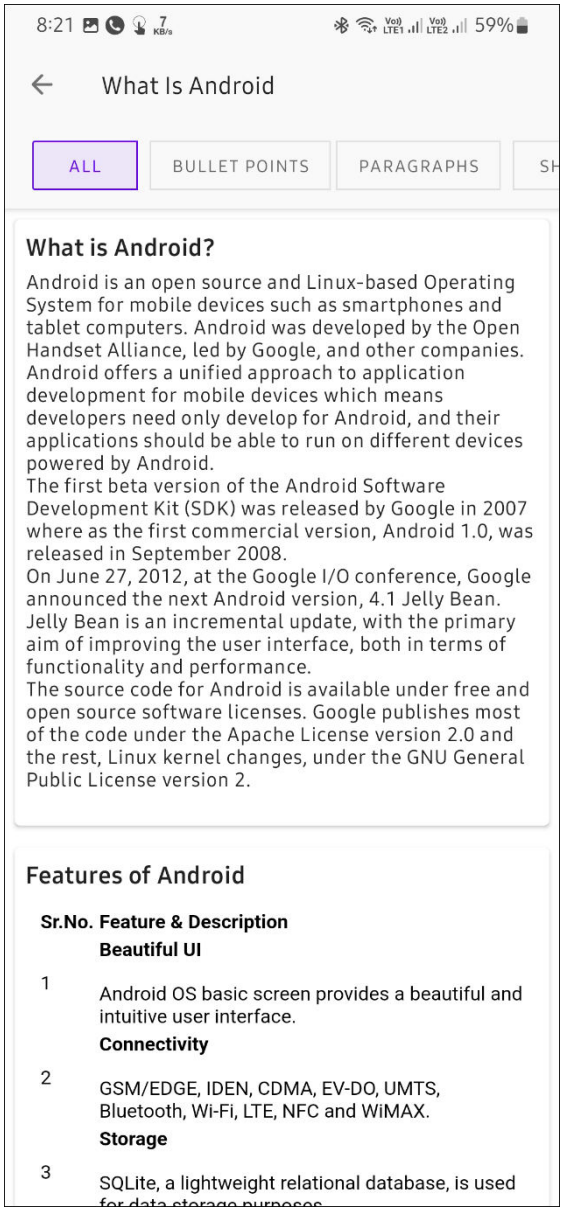

Fig. 3. Content Extraction Module

4. Database Module

The database module in this project is responsible for storing and retrieving website data obtained from the URL Visitor module. It organizes the website titles, URLs, and summaries in a structured format and provides a query interface for easy retrieval. The module ensures data integrity, security, and efficient management. It also serves as a data source for the presentation module, enabling the generation of visually appealing presentations with the website information. Overall, the database module plays a crucial role in storing, managing, and presenting the extracted

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

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205063 |

website data. Ex. INSERT INTO websites (title, url, summary) VALUES ('Example Website', 'https://www.example.com', 'This is summary of the website.');

5. Presentation Module

The presentation module in this project is responsible for retrieving data from the database and presenting it in a graphical layout. It retrieves the Title of the website, URL of the website, and the summary of the website from the database and displays them in the search results. Using a graphical layout, the presentation module organizes the search results in a user-friendly manner, making it easy for users to browse and select the desired websites. Each search result entry typically includes the website's title, clickable URL, and a concise summary that provides a brief overview of the website's content. This layout allows users to quickly scan through the search results and click on the relevant websites based on their interests. The user can even view the original web page without any filter by clicking the “visit original” button .

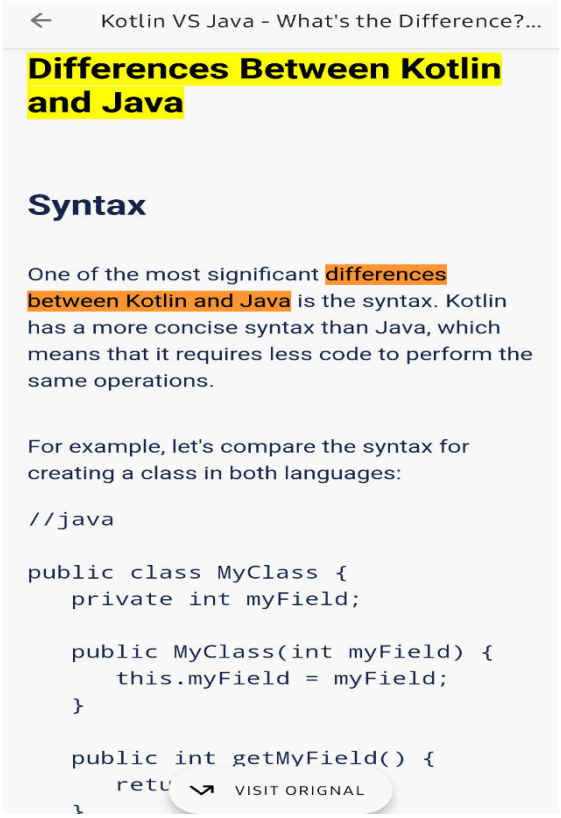

Fig. 4. Presentation Module

III. FUTURE SCOPE

The future scope of the project includes enhancing summarization techniques, providing user customization options, supporting multiple search engines, exploring cross-platform compatibility, integrating with social media platforms, implementing collaborative research features, enhancing privacy and security, optimizing performance, gathering user feedback, and integrating with external services like citation managers or research databases. These developments aim to improve the browsing experience, facilitate knowledge discovery, and enhance productivity for researchers

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

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205063 |

IV. CONCLUSION

In summary, the creation of an Android app with web scraping, clutter-free browsing, and summarising features gives considerable advantages to consumers and researchers looking for an uninhibited research experience. The project effectively built a functional software that allows users to conduct information searches, gather and store website data, and display summaries in a tidy manner. Future plans for the app include improvements to summarization methods, user customization, integration with more search engines and social media platforms, collaboration tools, privacy and security upgrades, performance optimisation, user feedback incorporation, and integration with external services. This initiative has the potential to revolutionise how users conduct research by giving them an effective and focused strategy to retrieve pertinent information while minimising distractions.

REFERENCES

[1]Koseeker Official. Koseeker. https://koseeker.com/careers, 2020. Online; Accessed 14 March 2020.
[2] Layla Hasan and Emad Abuelrub. Assessing the quality of web sites. Applied Computing and Informatics, 9, 01 2021.
[3] Yogesh Deshpande. Web Site Auditing – First Step Towards Re-engineering. Information Systems, pages 731–737, 2020.
[4] I. Sperano. Content audit for the assessment of digital information space: Definitions and exploratory typology. In Proceedings of the 35th ACM International Conference on the Design of Communication, SIGDOC ’17, New York, NY, USA, 2022. Association for Computing Machinery.
[5] K. M. Griffiths and H. Christensen. Quality of web-based information on treatment of depression: Cross sectional survey. British Medical Journal, 321(7275):1511–1515, dec 2022.
[6] Maria Manuela Cruz-Cunha and Joao˜ Varajao˜. E-business issues, challenges and opportunities for SMEs: Driving competitiveness. 2021.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details