Web Scraping Job Portals

Ashutosh Kumar, Kinshuk Chauhan, Jaspreet Kaur Grewal

Chandigarh University Mohali, Punjab-140413, India

Corresponding author: Kinshuk Chauhan, Email: ckinshuk01@gmail.com

Almost two million engineers and another million graduates in related areas like computer science and biotechnology are delivered in India each year. It can be challenging for work searchers to discover employments that fit their interface and aptitude level. The issues emerge from a need of information approximately the goals and operations of the organization. Web scraping of job portals gives insight into the most sought-after talents by recruiting organizations via the online job market, the industries that provide more work chances to job searchers, and other influencing elements to gain jobs, such as the candidates' experience. Scouts and work searchers might meet on a business entrance with the point of satisfying each other's person needs. They are the quickest and slightest costly way to communicate, coming to a wide gathering of people with fair one tap, wherever they may be within the world. The program settles these issues and offers work searchers a user-friendly stage for work looks and applications. Candidates can search for work in any field by utilizing progressed look methods.

Keywords: Web Scrapping, Employment portal, Job Market, SMTP, Authentication, IT Industry.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 291–303. DOI: https://doi.org/10.56155/978-81-955020-7-3-25

1. Introduction

1.1 Overview of Project

Web scraping is the process of obtaining necessary data or information from websites utilizing tools like Octa-parse, Parse-hub, and the Python computer language. Additionally, the gathered data must be transformed into a format that may be used for additional analysis. Searching for a job can be difficult whether you're reentering the workforce after a sabbatical or changing careers. As referred to Figure 1. all the necessary steps in the web scraping are shown briefly.

Job fairs, career administrations advertised by colleges, worker referrals, daily paper and T.V. advertisements, etc. are illustrations of conventional contracting strategies. With the development of technology and the rise in internet usage, e- recruitment has completely changed how companies make hiring decisions and how job seekers look for opportunities. The hiring process is expedited at every turn by using online job search portals, from advertising job openings to receiving applications from applicants to conducting interviews. Comparing the cost of job searching and posting to the conventional method of advertising will reveal significant savings. Employers may effectively attract job searchers by outlining job openings, responsibilities, and qualifications on job search portals.

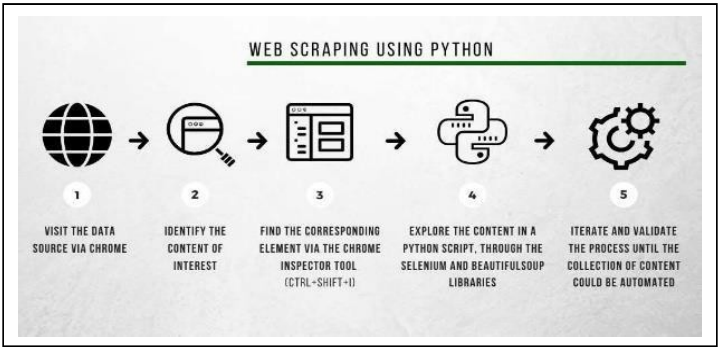


Figure 1. The General Mechanism of Web-Scraping. [1]

1.2 Importance of Profile Screening

The automated method of obtaining data from websites is known as web scraping. It has grown in significance in today's data-driven society for several reasons. With web scraping,you may easily collect enormous volumes of data from the internet. Numerous uses for this data exist, including data analysis and reporting, competitive analysis, and market research. Web scraping is a useful tool for businesses to keep an eye on their rivals and to check product offers, prices, and customer feedback.

1.3 Objective of the Research

The primary objective of this research project is to design, implement and evaluate a Platform for the people to get interest- oriented job applications profile's notification well suited from various Job-Profile Websites. As reference to the Fig 2. Our project seeks to address the following key goals by obtaining a structured data:

a. Searching suitable job-profile related to the description of skills of an individual.
b. Visiting correct and authentic websites for the search of jobs.
c. Reling the correct data to the right person with authenticity.

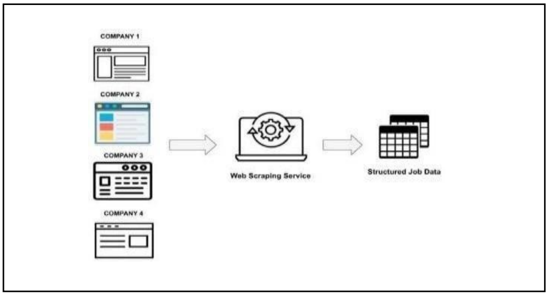


Figure 2. Web Scraping process to extract jobs. [5]

2. Literature Review

2.1 Introduction

A literature review is an outline of already distributed works on a specific subject. In Table 1. all the use cases of python is shown in details. The word may refer to a whole academic work or a segment of an insightful work, like a book or exposition.

a. **Python in Web:** Python is a powerful computer language that can be used to create websites. Python has a number of frameworks that make it easier to create web apps. Django, Flask, and Pyramid are among prominent Python web frameworks.

Table 1. Use of Python Technology

Use Case	Description
Web Development	Python is widely used for web development with frameworks like Django, Flask and Pyramid providing tools and features for building dynamic websites and web applications
Data Science	Python's extensive libraries, such as NumPy, Pandas, and SciPy, make it a popular choice for data analysis, machine learning, and scientific computing.
Artificial Intelligence	Python is widely used in AI development with libraries like TensorFlow and PyTorch providing powerful tools for building and training neural networks.
Automation and Scripting	Python's simplicity and versatility make it ideal for automating repetitive task and writing scripts for various purposes, such as systems administration and data processing.
Web Scraping	Python's libraries, such as BeautifulSoup and Scrapy, make it easy to extract data from websites, enabling task like data collection and content aggregation.
DevOps and Infrastructure	Python is used in DevOps for task like configuration management, development automation and infrastructure provisioning. Tools like Ansible and Fabric are popular in this domain.
Internet of Things (IoT)	Python is used in IoT projects for device communication and data processing and building IoT applications. Libraries like Raspberry Pi GPIO and PySerial are commonly used.
Game Development	Python has libraries like Pygame that simplify game development making it a popular choice for creating 2D games and protopypes.
Desktop Applications	Python can be used to build cross-platform desktop application using frameworks like PyQt and Tkinter.

b. **Python Libraries/Frameworks:** A library is a collection of code that improves the efficiency of routine tasks. In Table 2 the library and the framework of python is mentioned. Python frameworks automate the execution of many operations and provide developers with a foundation for application development. Each framework has its own set of modules or packages that considerably shorten development time.

Table 2. Some Common Libraries/Frameworks

Library/Framework	Description
Request	Library for making HTTP request and handling responses.
BeautifulSoup	Library for parsing HTML and XML documents.
Scrapy	Powerful web scraping framework for extracting data from websites.
Flask	Lightweight web framework for building web applications.
Django	High-level web framework for building complex web applications.
NumPy	Library for scientific computing with support for multi dimensional arrays.

2.2 Tools and Technologies Used in Project

2.2.1 Beautiful Soup

Figure 3. is a descriptive image of the package Beautiful Soup which makes it easier to extract data from html and XML files by turning them into a python object tree that can be navigated [2].

Figure 3. A Python web scraping package called Beautiful Soup

2.2.2 Selenium

Figure 4. Is Selenium tool which is well known for its ability to automate web applications.

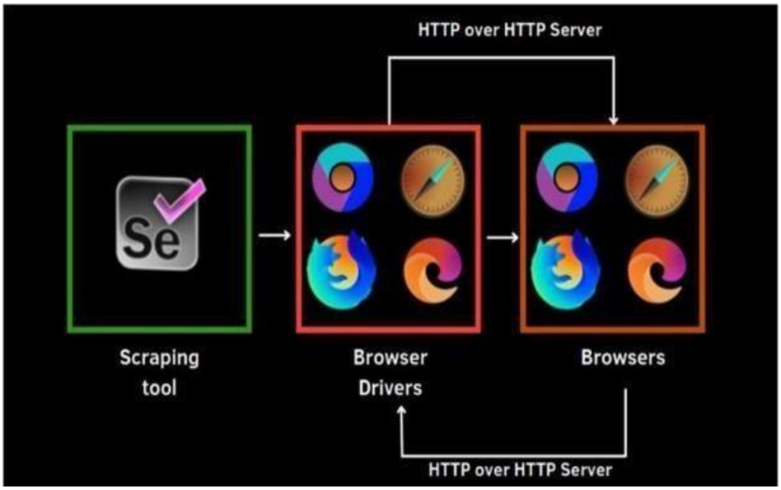


Figure 4. Selenium, a flexible open-source technology, usually used to automate the web pages build via html.

2.2.3 SCRAPY

Figure 5. Is a SCRAPY which is a crawling tool used a get data from the original web servers.

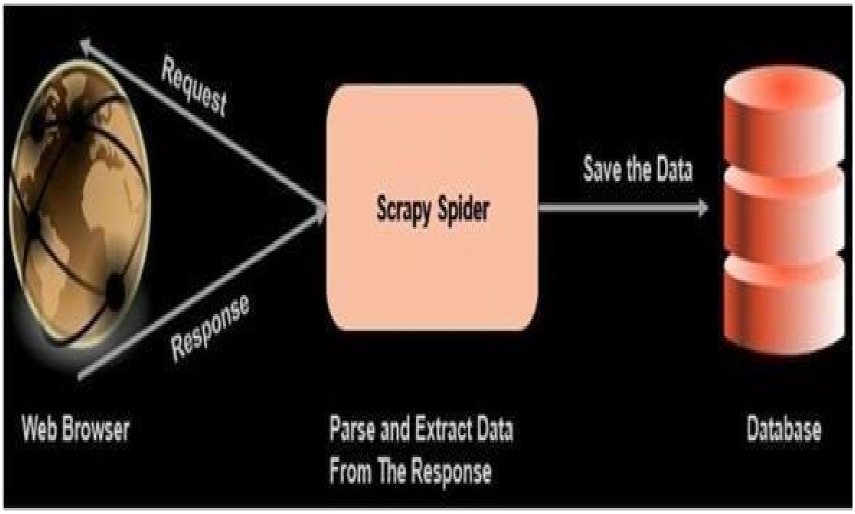


Figure 5. With the help of the robust open-source web crawling technology Scrapy.

2.2.4 JSON

In Figure 6. There is JSON is a common text-based format which is based on the JavaScript object syntax, is utilized broadly in web applications to encode structured data.

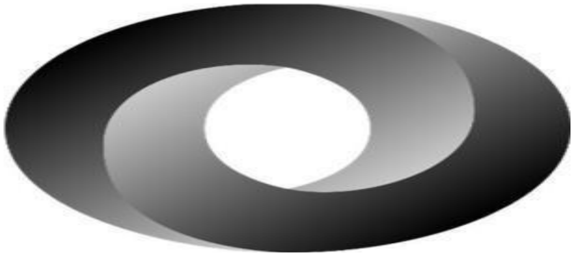


Figure 6. JSON is a lightweight, text-based format for exchanging data

2.2.5 API

In Figure 7. It is shown that we are going to use an API for the interaction of the web portal with the other server web portal.

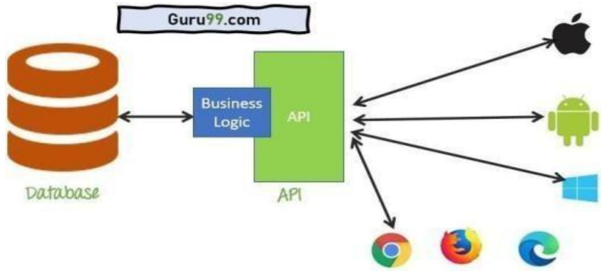


Figure 7. A user interface is in contrast to an application programming interface. It allows two or more computer programs to interact with one another.

3. Methodology

A clear strategy for evaluating the prerequisites of the Indian IT industry is to scrutinize work postings and get information from work portrayals. The larger part of person websites for Indian IT companies has a same plan for work promotions, which makes it simple for a computerized script to visit and assess each site on a customary premise. The work title, work portrayal, work encounter, and post-year are all included within the standard work posting fashion utilized by the larger part of firms. A few instances incorporate encourage subtle elements like instructive foundation, capacities, work space, work id/number, and business close date.

The top ten businesses in the Indian IT sector by market capitalization like Accenture, Capgemini, GEP India, Infosys, TCS and others are chosen, to perform this study. These corporations also hire a large number of undergraduates from Indian engineering institutes. Every firm's job search websites were first selected manually. We gathered employment data from several firms using web scraping.
Python's Selenium and Beautiful Soup modules provide assistant for web scraping. We extricated particular work joins from each company's work look site and completely inspected each work page to seek for data around the position, counting the title, postdate, area, and depiction, as well as

capabilities and involvement. It was ensured that there are no copy records within the dataset when including unused work information.

Step 1
Entering the URL of a website and sending an HTTP request. For instance, the server replies to the request by providing the HTML content of the webpage, for an instance
- Fresher Jobs Portal. A python library named ‘requests’ has been used to fulfill this work.

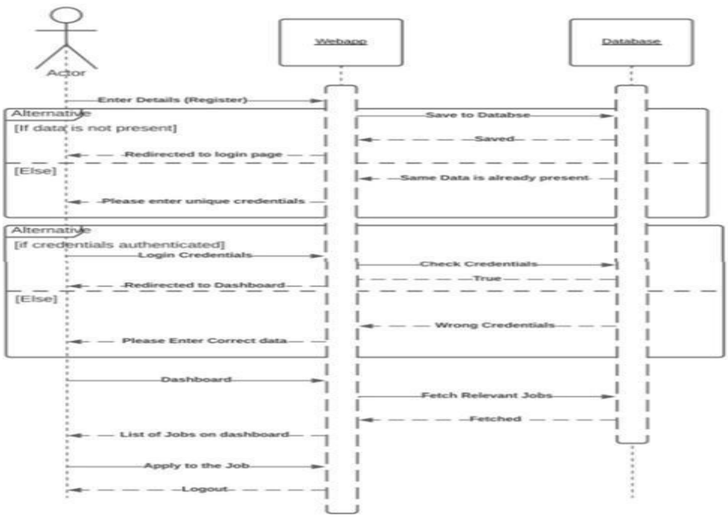
Step 2
After getting the whole HTML substance of the page, the information is parsed. Information cannot be extricated from HTML pages utilizing string handling on it possess since most HTML information is settled. It is essential to have a parser that can organize HTML information into a tree or settled structure.

Step 3
At this point, tree traversal—or browsing and searching the parse tree created—is all that is required. Beautiful Soup is utilized, another third-party Python module, for this purpose. This Python package is used to extract data from XML and HTML documents [3].

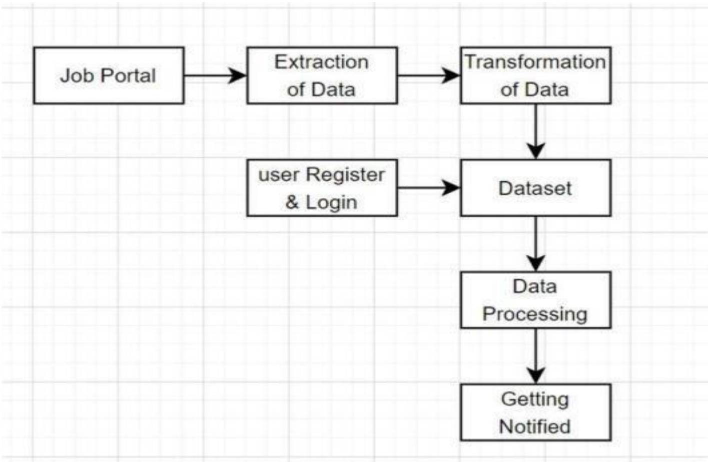
Step 4
The data will be kept in our MondoDB database when it has been extracted. It will be retrieved based on the needs of the logged-in user.

Step 5
Email alerts have been sent to the user based on the jobs that are most appropriate for them, using the SMTP protocol. The user will receive emails from our system twice daily [4].

a. Sequence Diagram



b. Data Flow Diagram



In today's society, there are several employment portals that provide career chances to job searchers. As reference to Table 3. It's a brief description of the libraries used in this project. The job searcher or the user have to register themselves on all job sites to find the finest- suited work for them based on numerous characteristics, such as e-mail address, name, Aadhar number, cellphone number, and so on. But, registering on every single site will be very hectic and time-consuming task as it takes minimum 10min to register on a single site. In order to register for our employment site, users must submit just basic information such as their Full Name, Mobile Number, Password (for your profile), Your talents (by selecting from a drop-down box), box of locations, and List of job titles. The user will view several employment alternatives from many job portals in one location after finishing the registration procedure. In consequence, our portal shortens the time needed for each employment portal's drawn-out registration process [6].

Table 3. Libraries/Frameworks and its Functions

Library/Framework	Capabilities/Functions
BeautifulSoup	HTML/XML Parsing data extraction
Selenium	Web browser automation interaction with web elements
Scrapy	Web scraping framework
Playwright	Web automation browser control

An additional analysis on the employment dataset, could help in determining the ranking of abilities depending on geography. This will assist Indian institutions in developing a curriculum that is tailored to certain regions. Knowing and extraction of the top 5 employment locations for specific talents from our dataset could lead to major improvements. Many corporations visit Indian institutions in an attempt to hire students, but the number of job offer letters that students receive is still very low since employers unable to find students with the adequate knowledge. The institution might enforce students to improve or update their abilities for a certain group of firms by bringing in top corporations for a particular expertise. It is been observed that there is a greater general need for students with Python proficiency at TCS and Mphasis (only intended as an example). Universities might adjust their course content based on trends to better align with the Indian IT sector.

4. Web Scraping Software

The features, support, and upgrade duration of web scraping software all are a major aspect for its pricing. To make sure it possesses all the functionality needed for scraping, one must make sure to always download the trial edition first.
There are various web scrapping tools available in the market that enables one to extract data from any website. Some of them are listed below:

4.1 Scrapy

Python-based Scrapy is an open-source, free web crawling framework created in Cambuslang. Online scrapping in not the only use of scrapy but can also be utilized as an API to retrieve data or as a web crawler. Its maintenance is currently handled by Zyte, a web-scraping development and services company. It is a high-level, quick framework for crawling websites and pulling structured data out of their pages.

4.2 Fminer

FMiner is a program that supports web macros and is used for scraping, harvesting, crawling, and extracting data from web applications. Other than that, it is also used for screen scraping, Windows and Mac OSX. It is a user-friendly web data extraction tool that makes your next data mining project a breeze by combining best-in-class functionality with a clear visual project planning tool [7].

5. Enhancing User Experience

An online portal that uses web scraping to gather job listings may greatly improve the entire experience of job searchers with an intuitive and effective interface.

5.1 Simple Registration Process

Rapid Enrollment: Reduction in the complexity of the registration process by simply requesting the most basic details such as name, phone number, and strong password. Getting rid of on obstacles for new users. Table 4 shows the importance of UI while customer open the portal.

Table 4. Principle of User Interface Design

Principle	Description
Simplicity	The user interface follows a minimalist design, promoting clarity and ease of use.
Accessibility	Accessibility features, such as voice commands and screen reader compatibility are integrated.
Visual Feedback	Real time visual feedback keeps users informed about device status and the outcomes of their actions.
Customization	User have the flexibility to customize the interface to suit their preferences and requirements.
Responsive Design	The interface adapts to various screen size and devices, providing a consistent user experience.

5.2 Customizing a User Profile

Prowess and Preferences: Gives preference to users for editing their profiles by letting them add or remove job titles, desired work locations, and skills. The employment recommendations are customized to meet their demands thanks to this customization.

5.3 Real-Time Updates

Automated Scraping: Use routine online scraping to maintain the job listings current so that users may see the newest openings as soon as they become available.

5.4 Email Notifications

Customized Alerts: Permit users to be notified via email when new job postings align with their interests. In addition to saving customers time, these alerts make sure they don't pass on pertinent possibilities.

5.5 Simple Search and Filters

Advanced Search: Make a search interface that is easy to use, including filters for talents, locations, job categories, and more. This makes it easier for people to locate employment that match their credentials and interests.

5.6 Detailed Job Listings

Detailed Information: Ensure that all important facts, such as the firm name, location, job description, skills required, qualifications, and application dates, are included in job ads. This data assists users in making educated judgments.

5.7 Mechanism for User Feedback

Feedback Channels: Create channels for users to submit feedback on the usability and functionality of the site. Utilize user feedback to continuously enhance the platform.

5.8 Performance Optimization

Page Load Speed: Make certain that the portal loads swiftly and effectively. Slow-loading pages can irritate users, resulting in a negative experience.

5.9 User Assistance and Resources

Help Desk: Provide tools like FAQs, tutorials, and guides to help users use the site successfully. Respond to user concerns and difficulties by providing responsive customer service. In Table 5. the importance of the data privacy and some important measures is shown.

5.10 Data Protection

Put in place strong security mechanisms to protect user data. Assure users that their personal data is secure.

Table 5. Measures To Ensure Data Privacy

Measure	Description
Data Encryption	Encrypt sensitive data for protection
Access Control	Restrict data access based on user role.
Data Minimization	Collect and retain only necessary data.
Anonymization	Remove or encrypt personally identifiable information
Consent Management	Obtain user consent for data collection.
Data Retention Policies	Establish policies for data retention and deletion.
Regular Security Audits	Conduct audits to identify vulnerabilities.
Employee Training	Train employees on data privacy best practices.

Incident Response Plan	Develop a plan to respond to data breaches.
Compliance with Regulations	Ensure compliance with data protection regulations

5.11 Keeping Data Private and Secure

It is critical to ensure the privacy and security of user data on a job site that uses web scraping. When enrolling and utilizing the site, job seekers provide sensitive information, and it is the administrators' obligation to keep this information secure. Table 6 shows some of the protocols for the encryptions.

Table 6. Security Protocols and Encryption

Protocol/Encryption	Description
SSL/TLS	Encrypt data during transmission
HTTPS	Secure connection between user's browser and job portal
ZFA	Additional verification for enhanced security
SHA	Ensures data integrity and password security
PKI	Public and private key pairs for secure communication
VPN	Secure connection between user's device and job portal
DES	Symmetric encryption algorithm for data encryption
AES	Widely used symmetric encryption algorithm
SFTP	Secure file transfer with data encryption
SSH	Secure remote access and file transfer

5.12 Access Control and Authorization

Access control and authorization frameworks are vital parts for overseeing user interactions on a web portal that scrapes job listings. Table 7 gives a insight about the access management and authorization rules of the portal .These safeguards protect both user data and the platform's operation.

Table 7. Access Control And Authorization

Measure	Description
Role-Based Access Control (RBAC)	Assigns access permissions based on user roles.
User Authentication	Verifies user identity before granting access
Access Logs	Records user access activities for auditing purposes
Access Restrictions	Implements restrictions based on IP addresses or time of access
Multi-Factor Authentication (MFA)	Requires multiple forms of verification for access.

By utilizing these attributes and tactics, your online portal may create a more gratifying and effective journey for job searchers, assisting them in more efficiently and successfully finding their perfect career prospects [8].

6. Results and Findings

This section illustrates everything that we as researcher(s) gathered through data analysis generally concerned with functionalities for improving the user experience. Even if the results defy the hypothesis, its main objective is to use the collected data to address the research question(s) stated in the introduction.

6.1 User Experience Enhancements

For a job site that uses web scraping to gather job listings, creating an amazing user experience is critical. A well- designed and user-friendly platform not only attracts but also retains users' interest and satisfaction.

6.2 Personalized Recommendations

Harnessing user data and activity to deliver targeted job recommendations. Use machine learning methods to ensure precise match.

6.3 Overall Assessments

The overall assessments of user experience and delivery of expectation and capabilities were overwhelmingly positive.

6.4 Discussion

The creation and deployment of a job seeker online portal that uses web scraping technologies to gather job postings is a big step toward expediting the job search process. This section delves into the research's major findings, their ramifications, and potential future directions.

a. **Aggregation of Data and Real-Time Updates:** The introduction of web scraping into the employment site has allowed for the aggregation of job postings from many sources, providing users with a more wide and diversified pool of options. Daily real-time updates guarantee that users have access to the most recent job postings.

b. **User-Centric Design and Personalization:** Improving user experience through features such as expedited registration, individualized user profiles, and targeted email alerts is critical to attracting and maintaining users.

c. **Data Privacy and Security:** The topic of data privacy and security emphasizes the company's commitment to protecting customer data. It is critical to foster user trust by establishing effective security measures such as encryption, access limits.

d. **Future Opportunities and Possible Expansion:** The efficiency of web scraping technology in the context of a job site was highlighted in this research report. However, there is significant room for growth and diversity.

e. **User Feedback and continued Improvement:** An active consumer input system is essential for the portal's continued improvement. As users submit feedback and recommendations, the site may adapt to changing user demands, guaranteeing its relevance in a competitive employment market.

f. **Responsible Considerations:** Because the portal interacts with third-party data sources, emphasizing the necessity of ethical data usage is critical. Respect for data ownership rights and compliance with legal agreements are key components of the portal's long-term viability.

7. Conclusion

Job Search Portals are an enormous portion of the recruiting industry. They fulfill the wants of both selection representatives and candidates by acting as a contact between them. This software helps employers by making them more obvious to the candidate pool, and it helps job seekers by making it simpler for them to conduct a careful seek for parts that fit their criteria. Without requiring to bring a portable workstation, job seekers can see accessible openings and yield applications utilizing the android application. The utilize of this application benefits bosses as well as job seekers. It can reach a wide group of audience because of its user-friendly UI. Each measure that was set forward amid the necessities gathering stage has been fulfilled by the application.

8. Acknowledgments

We would like to extend our whole hearted thanks towards our Project Guide/Supervisor Prof. Jaspreet Kaur Grewal, for believing in us, our talents, and keeping her trust in us that we would be able to accomplish the project on time, leading us through all crucial choices, and most importantly, believing in our abilities. This project was completed successfully thanks to her continual direction, support, and insightful input. Above all, thank you God for all of the chances you have presented us with.

References

[1] Ibef. Maio, Tbnsilveira (2020). Data acquisition, web scraping, and the KDD process: a practical study with COVID-19 data in Brazil | tbnsilveira.info

[2] L. Richardson, “Beautiful soup,” Jan 2020. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/

[3] S. d. S. Sirisuriya (November 2015). “A comparative study on web scraping,” 8th International Research Conference, KDU.

[4] S. Munzert, C. Rubba, P. Meißner, and D. Nyhuis (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons.

[5] Priam Pillai, Dhiraj Amin. Understanding the requirements Of the Indian IT industry using web scrapping. 9th World Engineering Education Forum, WEEF 2019.

[6] J. Ward, Instant PHP web scraping. PacktPublishing Ltd, 2013.

[7] F. Suleman, “The employability skills of higher education graduates: insights into conceptual frameworks and methodological options

[8] A. Radermacher and G. Walia, “Gaps between industry expectations and the abilities”.