Dataset Creation from Job Portals using Web Scraping to enhance the Employability

Prof. Priyanka Shah[1], Dr. Paresh V. Virparia[2], Dr. Hardik B. Pandit[3]
[1]S. K. Patel Institute of Management and Computer Studies, KadiSarvaVishwavidyalaya University, Gandhinagar-382021, Gujarat, India
[2]Department of Computer Science, Sardar Patel University, VallabhVidyanagar, Gujarat-India
[3]Faculty of Computer Science and Applications, Charotar University of Science & Technology, Changa, Gujarat, India

ABSTRACT
The Computer Science is a highly sought-after field, offering some of the best job opportunities in the Information Technology sector through various job portals. Students often find that their skills in computing can open many doors in the rapidly evolving tech industry. There's a significant gap between the skills taught in educational settings and those demanded by the industry, largely due to the rapid evolution of technology. To address this, educational institutions need to frequently update their curricula collaborate with industry leaders to ensure students acquire relevant, up-to-date skills for the workforce. Thus, it's essential to analyze the skills that are currently in demand within the IT industry. Recently, a multitude of data including text, images, audio, and video has become accessible online, and extracting useful information from these diverse web contents is a key application of Web Content Mining. A job portal is an online platform where companies can post job listings, offering a quick, reliable, and precise way to connect with potential employees. Web scraping has evolved into a crucial tool for obtaining important data from a variety of sources like Job portals. In this Research Paper, the Web Scrapping procedures and methods have been developed in Python to scrape Information Technology employment details from renowned job portals like Naukri.com, Monster.com, TimesJobs.com, Internshala.com and Myamcat.com.

Keywords: Web Content Mining, Web Scraping, Job Portal, Data Preprocessing, Employability, Key skills

1 INTRODUCTION

The World Wide Web (WWW) is a huge collection of diverse documents comprising text, images, audio, and video data [1]. Extracting data from websites has become a vital process known as web scraping. It involves retrieving information from multiple web sources, parsing the HTML content, and structuring it in a database for further analysis. The applications of web scraping encompass market research, competitor analysis, content aggregation, price monitoring, and more [2].

Python has emerged as the programming language for web scraping, owing to its simplicity, usability, and the availability of powerful libraries and frameworks tailored for this purpose. Notably, Python offers an wide range of libraries, such as BeautifulSoup, Scrapy, Selenium, and others. These libraries provide comprehensive features, including HTML parsing and

seamless interaction with web pages, resulting in streamlined and efficient web scraping processes [3].

In the context of this study, our primary focus is given to web scraping job portals, especially those that are related to the Information Technology (IT) sector. During Financial year 2022, The Indian Information Technology and Business Process Management Sector has gradually grown more than 30 percent of the global outsourced BPM market. In 2022, IT sector's contribution to India's GDP was 7.4%.[4] The IT industry is a significant source of employment and creates a wide range of job opportunities. However, there is a gap between the skills that job seekers possess and what the IT industry requires of them. Having access to pertinent and comprehensive datasets is necessary for identifying and closing this gap.

Despite the fact that a number of sources offer job data, none of them provide a comprehensive list of all IT job categories. Consequently, a specialized dataset becomes necessary. In order to specifically serve the IT industry, this paper addresses this need by utilizing web scraping techniques to gather job data from various job portals. Python's strong libraries make it possible to quickly find a variety of job listings .

2 PREVIOUS WORK

Author	Focus	Methodology	Finding
Mirjana Pejic-Bacha et al. (July 2019) [5]	The main objective of the research is the study of job advertisements in the field of Industry 4.0 using text mining techniques to identify the required knowledge and skills for Industry 4.0 jobs.	The research paper employs text mining techniques and utilize descriptive analysis, cluster analysis, and similarity measures such as Jaccard's coefficient to identify associations between phrases and determine key topics and skills associated with Industry 4.0 job profiles.	The study highlights critical job roles and skills essential for Industry 4.0, including Supply Chain Analyst and Digital Manufacturing Engineer. Key competencies like advanced analytics, digital transformation, and essential soft skills are emphasized, particularly for senior positions.
Francisco J. García-Peñalvoet al. (February 2018) [6]	The study explores the use of machine learning to assess employability, detailing the process from data collection to model creation, aiming to enhance the prediction of employment outcomes.	The study follows a machine learning-based approach. They use descriptive statistics, histograms, and bar charts to represent the information in a simple and understandable manner.	The study uses a machine learning model to predict employment outcomes with moderate success and identifies key employability factors, suggesting improvements in data processing to enhance future predictions.
Marianne P. Ang et al. (October 2017) [7]	The paper assesses how well the BS Information System Program in the Philippines aligns	The study uses qualitative methods to analyze how well the BS Information System Program's curriculum meets industry	The study at Ateneo de Naga University identifies that the BS Information System program aligns with 80% of industry

	with industry demands, highlighting the curriculum's effectiveness in meeting employer needs.	needs, finding an 80% alignment with professional skill requirements.	demands, recommending enhanced data analysis and syllabi updates to bridge the remaining skills gap.
David Smith et al. (2014) [8]	The study highlights the imperative for universities to align IT programs with industry needs, stressing the significance of teaching both modern and outdated technologies.	The methodology employed in the research involved web data mining techniques to collect and analyze a large volume of job data from job portals. The collected data was then analyzed to identify and interpret the trends in the job market.	The study highlights Java's market dominance and the decline of COBOL, emphasizing the value of SQL proficiency, especially with Oracle. It stresses the need for curriculum alignment with industry trends and effective data analysis methods to adequately prepare students for IT roles.
K.Thamarai Selvi et al.(January 2016) [9]	This study scrutinizes how electronic recruitment impacts job seekers, focusing on search engine and data mining tool usage.	The research methodology employed in this study involved data collection through Descriptive statistics, single sample tests, and one-way ANOVA and the utilization of the KNIME data mining tool.	This paper sheds light on the impact of electronic recruitment, emphasizing its effectiveness and accessibility for job seekers. It concludes that online methods are cost-effective and popular among diverse groups, facilitating easy access to job opportunities.
P Ramyaet al. (2020) [10]	This paper presents a comprehensive review of a study that focuses on the use of a recommendation system to enhance student performance through the application of machine learning algorithms.	The methodology outlines a comprehensive approach involving dataset creation, model development, and algorithm comparison for predictive analysis also emphasizes meticulous steps like pre-processing and separate training sets, ensuring robustness.	The conclusion emphasizes the transformative potential of Business Intelligence/Data Mining in education and resource optimization. It showcases the efficacy of leveraging past grades alongside socio-demographic factors for accurate predictions of student success in Mathematics.

Walid Shalaby et al. (2017) [11]	The paper discusses improving job recommendations on large online job boards by highlighting the shortcomings of content-based systems and stressing the need for more accurate matching methods.	The research introduces a graph-based approach to job recommendations that leverages deep learning and user behaviour to overcome scalability, sparsity, and the cold-start problem, providing personalized hybrid recommendations for new users and jobs.	The study demonstrates that a graph-based method, integrated with deep learning and user behaviour analysis, significantly improves job recommendation accuracy and coverage.

3 METHODOLOGY

Following Steps are required to create a Dataset for Information Technology Jobs from Job portals using Web Scrapping.

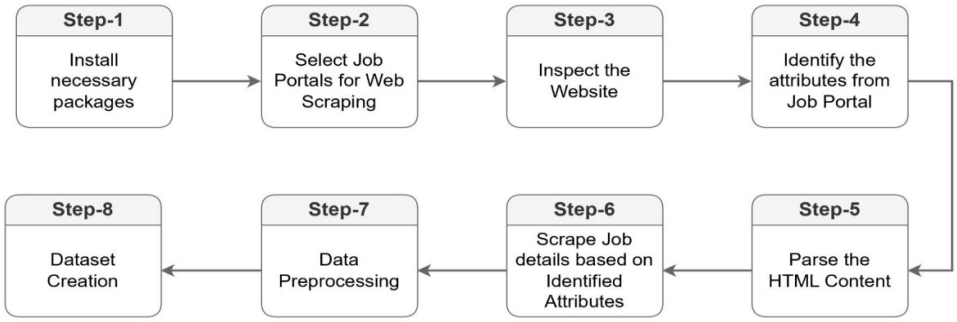

Fig. 1: Methodology to create a Dataset from Job Portals

Step 1: Install the Essential packages

The first step is to install the essential **Python libraries** and tools that will be used for web scraping. BeautifulSoup and Requests are the two most popular libraries used for web scraping in Python. While Requests is used to send HTTP requests to websites, BeautifulSoup is used to parse HTML and XML documents. The Python package manager pip can be used to install these libraries by executing the commands below in the terminal:

pip install beautifulsoup4
pip install requests

Step 2: Selecting Job Portals for Web Scraping

Numerous job portals such as Monster.com, Naukri.com, and Internshala.com provide platforms for employers to advertise job openings across various fields. The structure of a website can vary, and there are different techniques used to fetch data from it. Therefore, separate Python modules have been developed for each job portal, tailored to their specific website architecture.

Step 3: Inspect the website

In this step, the website's HTML structure must be examined to scrape the desired data. Some websites only provide static content, which consists of pre-rendered HTML pages that don't

include any dynamic information. If the website has a static structure then select the Inspect Element menu from web browser then move the cursor to the desired data. After that step, copy the CSS selector and check the CSS selector by chrome extension CSS/XPATH selector. After selecting the precise elements for scrapping, web scraping tool can be used or source code can be written to extract the desired data from the website. Other websites implement dynamic content delivery, which involves data retrieval via APIs or other methods like WebSockets, server-sent events, or network requests. Python can be used to build modules that extract data from websites that implement dynamic content delivery. So, In dynamic content delivery, the API network request is tracked and mimicked within a Python script to retrieve the data. in dynamic content delivery, the API network request is being monitored and the network request is being spoofed in a Python script to get the data. Most of the data sent by API is in the form of JSON from which specific attributes can be extracted. Fig.2 shows the steps to inspect the website's structure.

Some websites only provide static content, which consists of pre-rendered HTML pages that don't include any dynamic information. If the website has a static structure then select the Inspect Element menu from web browser then move the cursor to the desired data. After that step, copy the CSS selector and check the CSS selector by chrome extension CSS/XPATH selector.

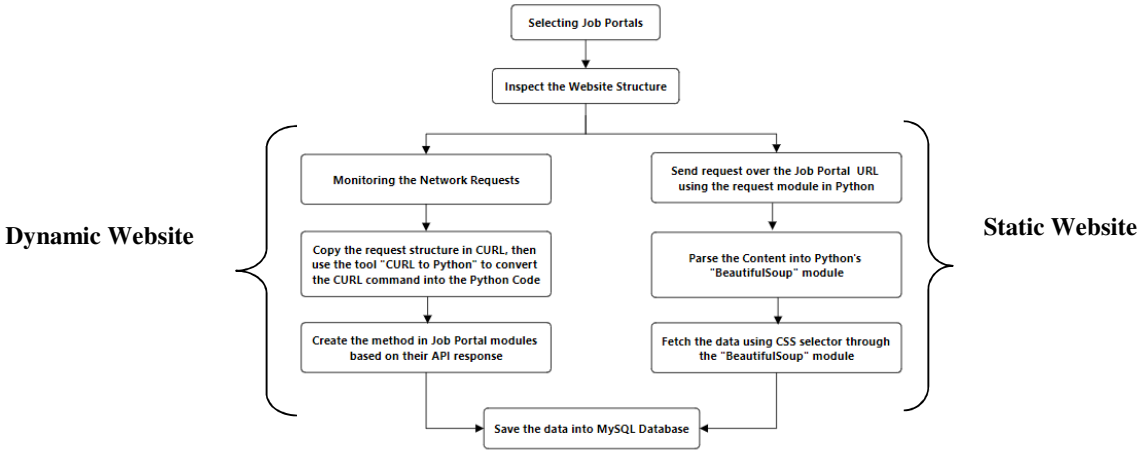

Fig. 2 Inspect the Website Structure

After selecting the precise elements for scrapping, web scraping tool can be used or source code can be written to extract the desired data from the website. Other websites implement dynamic content delivery, which involves data retrieval via APIs or other methods like WebSockets, server-sent events, or network requests. Python can be used to build modules that extract data from websites that implement dynamic content delivery. So, In dynamic content delivery, the API network request is tracked and mimicked within a Python script to retrieve the data. in dynamic content delivery, the API network request is being monitored and the network request is being spoofed in a Python script to get the data. Most of the data sent by API is in the form of JSON from which specific attributes can be extracted.

Step 4: Identify the attributes from Job Portals

At this stage, identify and select only those attributes from the job domain which are required. Identify the attributes from Job Portals like Job Title, Salary, Experience, Education, Key Skills, Location, JD, Source, JD_DATE, JB UNIQID.

Step 5: Parse the HTML or JSON Data

If the website is providing static content, then a parser must be used like BeautifulSoup to parse the website's HTML data.

To do this, make a BeautifulSoup object and pass the page's HTML content to it as follows:

```
import requests
from bs4 import BeautifulSoup
url = "https://jobportal.com"
response = requests.get(url)
soup = BeautifulSoup(response.content, "html.parser")
```

If the website is providing dynamic content then Python Request Library and parse the data in JSON form must be used.

```
import requests
url = "https://jobportal.com"
response = requests.get(url)
 data=response.json()
```

Step 6: Identify attributes and scrape Job Details

Web scraping involves pulling information from online sources and organizing it into a usable format. This step must precede data preprocessing and analysis. For this purpose, a scraper script was created in Python, and 72 specific keywords were established, as displayed in Table 1, to extract Information Technology job listings from various job portals.

Table 1: 72 Keywords

C++	Software Developer	Application	Web Developer	J2EE	DBA
C	Information Security	Database Manager	Machine Learning	Blockchain	AWS
Web	Content developer	Security	Game Designer	Android	Cloud
PHP	Network Security Engineer	Agile	Data Scientist	Javascript	IOT
JAVA	Computer System Analyst	Data Analyst	Graphics Designer	Angular	SAP
Flutter	Mobile Application	Programmer	Software Tester	.NET	Node JS
QA	Cyber Security Engineer	Content Writer	Content Writer	SQL	Back end
Testing	Database Administrator	Automation	IOS developer	Azure	Unix
Oracle	Embedded Engineer	ASP.NET	Dot net Developer	AEM	R
Python	System Programmer	UX Designer	Quality Assurance	DevOps	Laravel
Linux	React JS Developer	Ecommerce	Data Entry Jobs	Full Stack	API
C#	Angular JS Developer	SEO executive	SMO Executive	Front End	Database

Method creation for Job Portal Modules: Different modules and methods have been created for each job portal module to get the attributes because every job portal is differently structured. Separate methods have been for each job portal module to handle scraping tasks.

Each method should be customized to the specific structure and requirements of the job portal.

6.1 Scrape individual Pages: The "scrapeSinglePage" method have been implemented to scrape all the job links in a specific job page.

6.2 Navigate all Pages: The "scrapeAllPages" method have been developed to navigate through different pages within the job portal. Implement the method to jump from one page to another, following the navigation links or pagination.

6.3 Scrape Job Pages: The "scrapeJobPage" method is used to extract attribute data from individual job pages. The required attributes have been defined to scrape, such as job titles, company names, locations, or descriptions. Implement the logic to extract the desired data from the job page HTML structure or using their internal API.

6.4 Implementing Python Threading for Concurrent Execution: Separate threads have been created for each job portal module. And as the scraping tasks for different job portal modules run concurrently, this can improve efficiency by allowing multiple job portals to be scraped simultaneously.

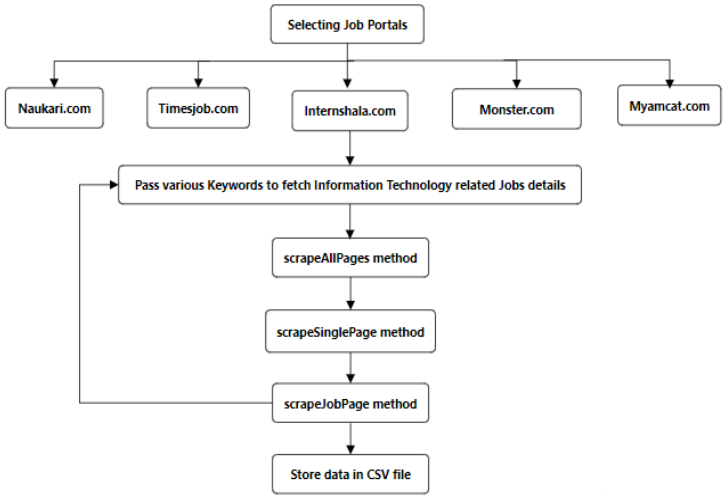

Fig. 3: Web Scraper

By utilizing separate methods for different scraping tasks, the scraping process can be modularized and make it easier to handle variations in website structure across different job portals. Implementing threading allows for concurrent execution, enabling efficient scraping of multiple job portals simultaneously, as shown in below Figure 3.

Step 7: Data Preprocessing

Data Preprocessing involves following steps:

7.1 Interpolation: The data extracted might lack cleanliness and could contain duplicates or missing values, necessitating the application of diverse techniques to rectify these issues and identify any missing data.

7.2 Feature Extraction: Not all the extracted data is pertinent for mining purposes, thus necessitating the selection of only the relevant attributes. Data preprocessing was then carried out on the obtained dataset, incorporating User-Defined Functions (UDFs) to clean the data.

This involved removing missing values, duplicates, as well as any extraneous characters such as new lines and tabs from the strings, ensuring the dataset was free from anomalies.

4 RESULT

Multiple relevant websites have been explored using their URL addresses. Scrapper was generated and implemented using Python. The scrapped data set is retrieved in an CSV format. From the above methodology, total 10217 job data were scrapped as shown in figure 4.

Sl	Title	Salary	Experience	Education	KeySkills	Location	JD	Source
1	Senior Software Developer (JAVA	900000 - 900000 INR	Min. - 0 Max. - 10	Any Graduate in Any Spe	Linux | RDBMS | TDD | Postgresql | Analytical | Manager Technolo	Bengaluru	Â We arelookir	naukri.com
2	Associate Software Engineer	3.3 - 3.3 LPA	0 - 2 Years	B.Tech/B.E., M.Sc. (Tech.	CI | Environment | ich | IT | JAVA | Oracle | Python | Security | Tec	Nagpur	Please review the	myamcat.co
3	Social Entrepreneurship (Social Vc	Unpaid	Internship	NULL	Creative Writing |	Work From H		internshala.
4	Senior software developer - C / C#	800000 - 800000 INR	Min. - 0 Max. - 10	Any Graduate in Any Spe	Communication protocols | Lead Software | Linux | Analytical | Ar	Bengaluru	Â What part will	naukri.com
5	Jr. Software Developer	500000 - 500000 INR	Min. - 0 Max. - 3	B.Tech/B.E. in Computer	dotnet core | angular | dotnet | XML | JSON | mvc | Win Forms	Ahmedabad	â€¢ B.E in computer	naukri.com
6	Opening For Frehers # Software D	350000 - 350000 INR	Min. - 0 Max. - 0	Any Graduate in Any Spe	Software Services | Development | ITES	Bengaluru	Dear candidates,We	naukri.com
7	Junior Software Developer / Softv	600000 - 600000 INR	Min. - 0 Max. - 3	B.Tech/B.E. in Any Speci	Hibernate | software development | maven | spring cloud | Spring	Bengaluru	Paychex, Inc. (NAS	naukri.com
8	Jr Software Developer	500000 - 500000 INR	Min. - 0 Max. - 3	Any Graduate in Any Spe	Computer science | Object oriented design | Manager Quality Ass	Chennai	Â Whatpart will	naukri.com
9	Software Developer Trainee(java/	300000 - 300000 INR	Min. - 0 Max. - 1	BCA in Computers, B.Tec	c# | C++ | python | software development	Bengaluru	Still waiting to see	naukri.com
10	Software Developer (java/python	400000 - 400000 INR	Min. - 0 Max. - 2	B.Tech/B.E. in Any Speci	c# | python | C++	Chennai	Still waiting to see	naukri.com
11	Embedded Software Developer	400000 - 400000 INR	Min. - 0 Max. - 4	Any Graduate in Any Spe	Maven | C++ | Linux | Eclipse | Configuration management | SOC	Kharagpur	Requirement:	naukri.com

Fig. 4 : Scraped Data in CSV format

From Scraped Data, key skills are also separated and stored in CSV and SQL databases. This approach makes it simple to analyze and identify the skills that are in demand on the job market, which can help job seekers and educators decide which skills to focus on acquiring.

5 CONCLUSION

In conclusion, the dataset generated through this web scraping method, focusing on job data from various job portals, holds significant potential for bridging the gap between academia and industry requirements. This study aims to use Python's advantages and its strong libraries to scrape data from job portals, with a focus on the IT sector. We aim to close the skill set gap between industry requirements and available skill sets by building a comprehensive dataset. By analyzing this dataset, researchers and educators can gain valuable insights into the skills that are in high demand in the job market. The availability of such a dataset enables a deeper understanding of the skills that are repeatedly emphasized in job listings, providing valuable information for educational institutions and students. By aligning their curriculum and skill development programs with the identified in-demand skills, educational institutions can better prepare students for industry-level job opportunities. The skills are also identified that employers are required the most through this analysis. The abilities are also determined that are in great demand for industry-level job opportunities by identifying the skills that are frequent in the scraped data. In Future enhancement, the knowledge provided by this web scraping method has the potential to support research, guide curriculum development, and improve student skill development.

REFERENCES

[1] A. Richlin, S. Jebakumari, and N. J. Goldena, "A Survey on Web Content Mining Methods and Applications for Perfect Catch Responses," *Int. Res. J. Eng. Technol.*, p. 407, 2008.

[2] Fatmasari, Y. N. Kunang, and S. D. Purnamasari, "Web Scraping Techniques to Collect Weather Data in South Sumatera," *Proc. 2018 Int. Conf. Electr. Eng. Comput. Sci. ICECOS 2018*, vol. 17, pp. 385–390, 2019, doi: 10.1109/ICECOS.2018.8605202.

[3] P. Thota and E. Ramez, "Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis," *ACM Int. Conf. Proceeding Ser.*, pp. 306–314, 2021, doi: 10.1145/3453892.3461333.

[4] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[5] M. Pejic-bach, T. Bertoncel, M. Meško, and Ž. Krstić, "International Journal of Information Management Text mining of industry 4 . 0 job advertisements," *Int. J. Inf. Manage.*, vol. 50, no. July, pp. 416–431, 2020.

[6] F. García-Peñalvo, J. Cruz-Benito, M. Martín-González, A. Vázquez-Ingelmo, J. C. Sánchez-Prieto, and R. Therón, "Proposing a Machine Learning Approach to Analyze and Predict Employment and its Factors," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 2, p. 39, 2018, doi: 10.9781/ijimai.2018.02.002.

[7] M. P. Ang, "EVALUATING THE RESPONSIVENESS OF BS IS CURRICULUM TO INDUSTRY SKILL REQUIREMENTS Aileen L . Rillon , MBA , MIM A Research Report Submitted to the Ateneo De Naga University Research Council October 2017," no. October, 2017.

[8] D. T Smith and A. Ali, "Analyzing Computer Programming Job Trend Using Web Data Mining," *Issues Informing Sci. Inf. Technol.*, vol. 11, pp. 203–214, 2014, doi: 10.28945/1989.

[9] K. T. Selvi and R. Ramya, "an Impact on Electronic-Recruitment and Its Perception Towards Job Portal Function Through Search Engines Among Job Seekers Using Knime Data Mining Tool," *Int. J. Res. Dev. Technol.*, no. 5, pp. 10–18, 2016.

[10] P. Ramya, S. G. Balakrishnan, and M. Kannan, "Recommendation system to improve students performance using machine learning," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 872, no. 1, 2020, doi: 10.1088/1757-899X/872/1/012038.

[11] W. Shalaby *et al.*, "Help me find a job: A graph-based approach for job recommendation at scale," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 1544–1553, 2017, doi: 10.1109/BigData.2017.8258088.