JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A STUDY ON WEB SCRAPING OF SELECTED JOB PORTALS

*L. GANGADHARA REDDY

** DR.P. VISWANATH

*** MBA – BDA** student, school of management studies, JNTUA, ANANTAPUR-5152002

**ASS. PROFESSOR (A), school of management studies, JNTUA, ANANTAPUR-5152002
(pvnmbajntua@gmail.com)

ABSTRACT:

The web scraping of job portals provides a view on the highly demanding skills by recruiting companies through online job market, the industries providing higher job opportunities to job seekers and the other influencing factors to get jobs like experience of the candidates. This study identifies the locations which are providing more job opportunities in India through selected job portals (Naukri, Indeed, LinkedIn). The results found that Business Management, IT-Programming skills(Java, Python, JDBC, SQL, Data Management, CSS) are having higher job requirements followed by selling skills. And IT-Software industry (41%) providing highest job vacancies followed by Banking(8%), Recruitment(8%) industries, among all other cities Bangalore(20%) providing most job opportunities. And the study presents top-20 job titles , top-20 Roles offered by recruiters to the job seekers.

INTRODUCTION

Web scraping is the process of collecting required information or data from web sites with help of python programming language or by using applications like Octaparse, Parsehub etc. And the collected data need to be converted into usable format to use for further analysis process. The web scraping can be performed on e-commerce sites to extract product related data, Job portals for collecting jobs related data, Entertainment, social media sites. Unlike the boring and dull process, mind-numbing process of manually extracting data, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet's seemingly endless frontier.

Web scraping has been around since the birth of the internet, but not many people seem to know about it. Ironically, the success of web scraping as a business tool has contributed to its under-the-radar-status; companies who rely on web scraping don't want competitors gaining access to their secret weapon. Automating tasks can go hand in hand with using the World Wide Web to connect with students and researchers, and obtaining information readily available to learn more about educational practices and preferences, as well as the dissemination of results.

MEANING:

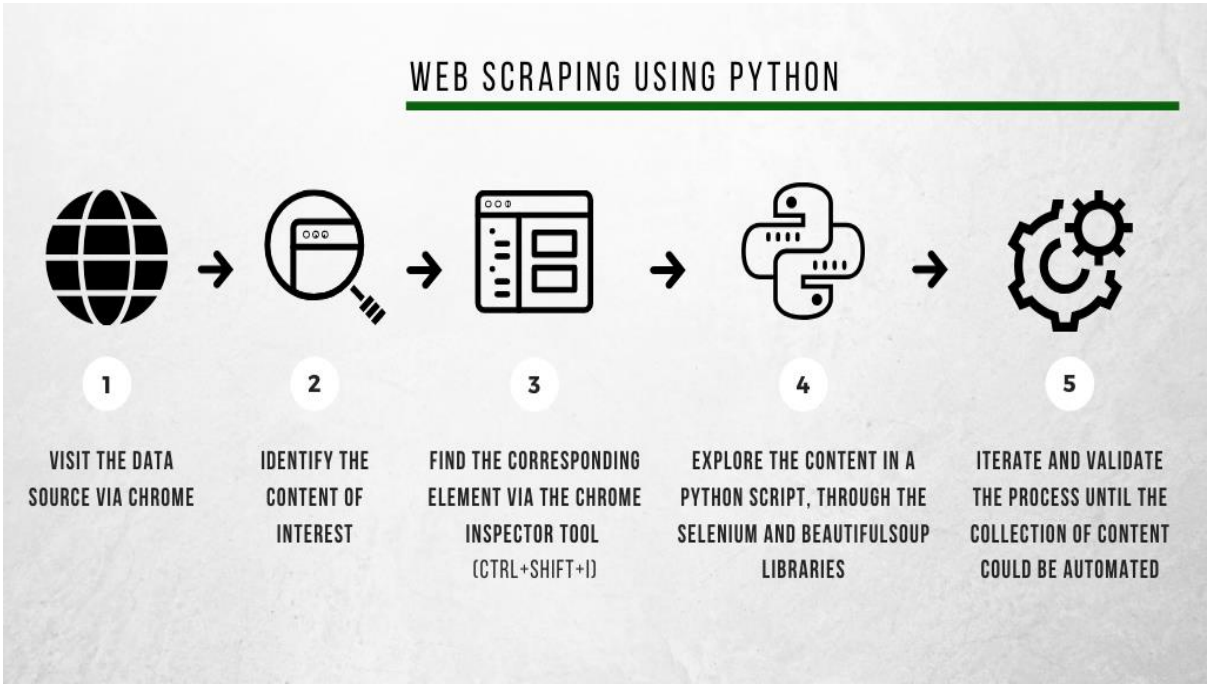
Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler.

WEB SCRAPING OF JOB PORTAL:

Web Scraping of job portals is the process of extracting or retrieving jobs related data from internet jobs publishing websites and transforming the collected unstructured data into structured and usable formats. In the study we are using usual combination of Request and Beautiful Soup, to capture the HTML content and then convert it to a Beautiful Soup object to parse through it easily and extract the data points.

The study examines the current trends in industry hiring of Job seekers (students). Implementing web scraping on Indeed.com for job postings, we create a structured dataset. Then, we apply data analytics, visualization techniques to identity trends including the major job titles, most in demand skills requested, vacancies by city, and degree preferences.

WEB SCRAPING PROCESS USING THE PYTHON LANGUAGE:



Source: https://tbnsilveira.info/2020/05/23/the-data-acquisition-process-via-web-scraping-a-case-study-of-covid-19-in-brazil/

Web scraping has been around since the birth of the internet, but not many people seem to know about it. Ironically, the success of web scraping as a business tool has contributed to its under-the-radar-status; companies who rely on web scraping don't want competitors gaining access to their secret weapon. Encountering the world of web scraping for the first time can feel like discovering a new and uncharted continent. Luckily, once you get your bearings, it's not too hard to navigate.

MARKET SHARE OF WEB SCRAPING:

In addition, an increase in research and development activities in different industries is expected to fuel the growth of the demand for web scraping software. With a total market share of around 50 percent, the e-commerce industry is the largest user of web data.

Global Web Scraper Software Market is expected to project a notable CAGR of 3.75% in 2030. Global Web Scraper Software Market to surpass USD 196.88 million by 2030 from USD 149.09 million in 2018 at a CAGR of 3.75% throughout the forecast period, i.e., 2019-30. The growth of the market for web scraping software is driven by key factors, such as development activity, in line with the current market situation and demand, market risks, new technology assessments, acquisitions, new trends and their implementation.

REVIEWE OF LITERATURE:

Renita Crystal Pereira et. al., provided web scraping summary and techniques and tools that face several complexities as data extraction isn't that simple. These strategies guarantee that the data collected is correct, consistent and has better integrity, because there is a large amount of data present which is hard to handle and retain. Although there are a few problems faced by functional techniques that can be such as the elevated amount of web scraping be able to cause rigid harm to the websites.

Kaushal Parikh et. al., proposed a web scraping detection with the help of machine learning It is valuable for research dependent companies. Web scraping has forever been a difficult preventive attack. Every time a company places its data on internet, it is probable that it could be copied and pasted and then utilized in the other point of view without the corporation knowing itself about it. A lot of protection mechanisms have already been in place but some of them continue to be ignored.

Sameer Padghan et. al., projected an approach where data extraction is done from web pages in assistance with web scraping easily. This method would enable the data to be scrapped from numerous websites that will minimize human intervention, save time and also enhance the quality of data relevance. It will also support the user in gathering data from the site and to save the data to their intent and use it as the individual wishes. The scraped information may be used for database development or for research purposes and also for different similar activities.

Anand Saurkar et. al., discovered latest technique named Web Scraping. Web scraping is a quite important methodology used to produce structured data based on the unstructured data available on the internet. Scraping formed structured data, subsequently collected and evaluated in spreadsheets in central database. This research focuses on a summary of the data extraction process of web scraping, various web scraping strategies and most of the latest tools utilized to scrap web. They concentrated on the Web scraping techniques.

Federico Polidoro et. al., concentrated on the outcomes of web scraping evaluation strategies with particular orientation to user electronics services and goods throughout the sector of commodity price studies. Although the research done has so far been performed in a small amount of time, that you can see in whatever followed, it has enabled to attain important, but not conclusive, novel efficiencies results..

Jan Kinne et. al., Proposed a web extraction platform for the accurate and measurable mining of ecosystems for development. Researchers have put special emphasis on exploring a possible bias while examining technology structures across corporation website if all those types of companies could be measured using suggested method. Web extraction still has to deal with incredibly large and ultra-connected outer websites as a research tool, and the reality that limited broadband access continues to discourage companies from managing their internal websites and therefore preventing themselves from web mining research.

Ingolf Boettcher discovered that technique like web scraping can evolve. Web scraping innovation provides a range of choices and can satisfy various purposes: A web crawler's ultimate requirement will be to discover previously inaccessible pricing data outlets and include a census of all web-available price information. The actions to build web scraping for price analytics include significant analytical and administrative consequences.

Erin Farley et. al., destined to present web scraping to law enforcement researchers and illustrate what web scraping is about and how this technique works. Use of the web crawling by investigators in criminal justice is a fairly recent trend. Although web scraping is usually seen as a method for collection of data to promote analysis and research, designing and implementing a web scraper includes technological abilities that researchers in the social sciences generally do not have. A strong level of expertise in computer science techniques like R or Python when developing source code is a necessity for creating a web scraper

On the whole the review of literature reveals that there is a gap. At this context extracting data and analysing under various dimensions in all areas like healthcare, social media, e-commerce sites benefits stake holders. Job portals extracting data on various factors which assist to know the most important skills required for both the job recruiters and job seekers. Hence the study is undertaken.

NEED OF THE STUDY:

The technological advancement in all sectors forced to upgrade skills for both business community and job seekers. At this context there is a need to study skills in demand for various job positions at different locations.

OBJECTIVES OF THE STUDY:

- To explore skills in demand in selected job portals.
- To study industry wise job vacancies in selected job portals.

SCOPE OF THE STUDY:

The study covers the extraction of data from selected job sites for period of two months i.e. (1-June-2022 to 31-July 2022).

RESEARCH METHODOLOGY:

The present study made on secondary data from Job Websites, Journals, Articles.

DATA TOOLS:

Python programming language has been used to extract the data and convert into csv file. Tables, frequencies have been used for descriptive analysis. Further, with help of Power BI charts, plots has been build to visualise the findings.

LIMITATIONS:

- The study is limited to selected job portals only.
- The study considers present available jobs posted in websites only.

DATA ANALYSIS AND INTERPRETATION:

For the data analysis and interpretation, skills in demand, industry wise vacancies and top job titles, roles of selected job portals has been analysed and interpreted.

ANALYSIS OF SKILL WISE JOB VACANCIES:

Broad classification of skills are Technical skills(53%), Soft skills(47%) among total skill requirements.

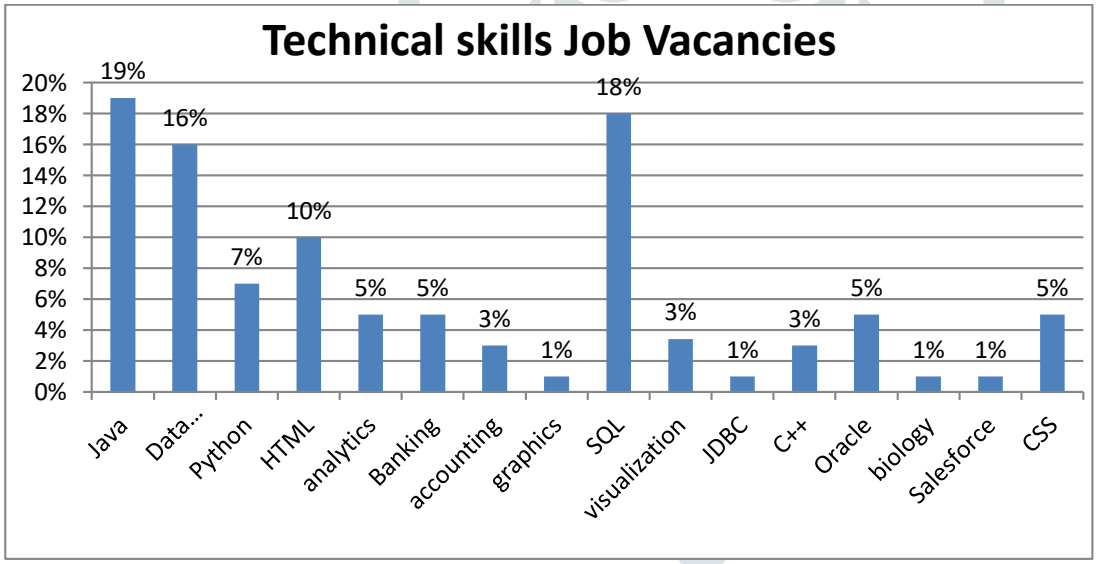
Skills	Vacancies	Percentage
Technical skills	17825	53%
Soft skills	15430	43%
Total Vacancies	33255	

In the study skill requirements for jobs classifies into Technical skills-java, data management, python, HTML, analytics, banking, accounting, graphics, SQL, JDBC, C++, oracle, biology, sales force, CSS. And in Soft skills namely interpersonal skills, management skills, communication, administration, predictive and selling skills.

TABLE 1: SKILL WISE JOB VACANCIES:

Technical Skills	Percentage	No.of jobs	Soft Skills	percentage	No.of jobs
Java	19%	3475	interpersonal skills	2%	339
Data Management	16%	2835	Business Management	46%	7055
Python	7%	1186	communication	8%	1266
HTML	10%	1856	Administration	6%	865
analytics	5%	831	Predictive Modeling	0%	63
Banking	5%	818	Sales	32%	4994
accounting	3%	498			
graphics	1%	233			
SQL	18%	3259			
visualization	0%	607			
JDBC	1%	101			
C++	3%	510			
Oracle	5%	911			
biology	1%	97			
Salesforce	1%	206			
CSS	5%	942			
		17825			15430

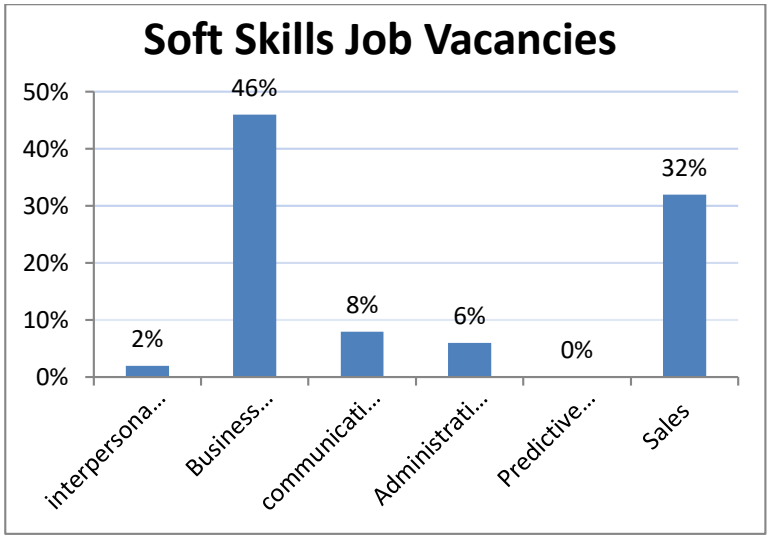
CHART 1: TECHNICAL SKILL WISE JOB VACANCIES:



INTERPRETATION:

The chart 1 reveals that in the technical skills the highly demanded skills are Java(19%), SQL(18%), and Data Management(16%).

CHART 2: SOFT SKILLS WISE JOB VACANCIES:



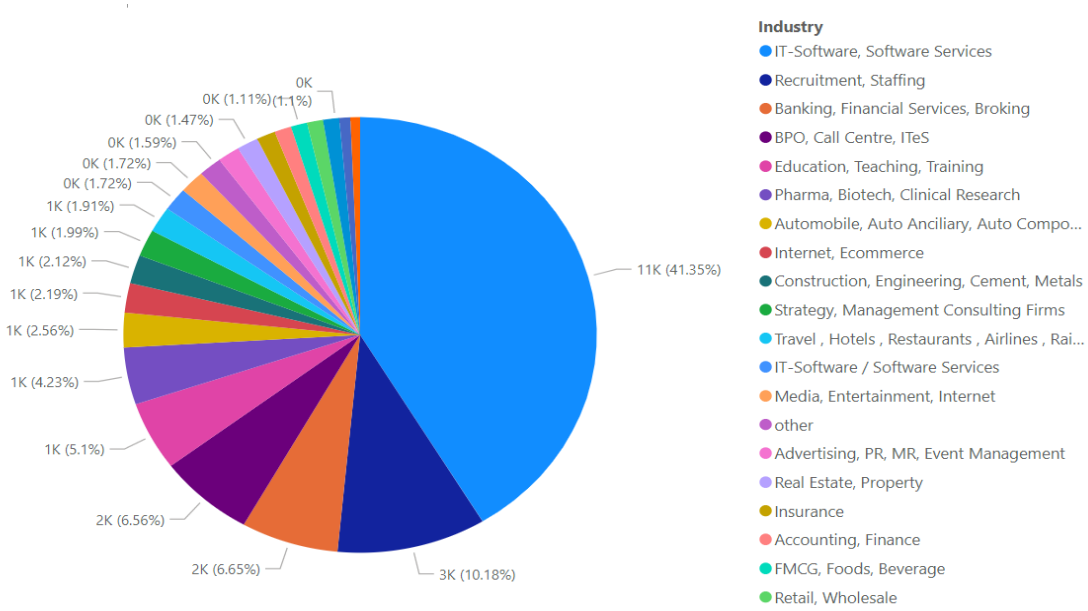
INTERPRETATION:

The chart 2 reveals that in the Soft skills the highly demanded skills are Business Management(46%) and Selling Skills(32%).

TABLE 2: INDUSTRY WISE JOB VACANCCIES:

Industry	vacancies	Percentage
Advertising, PR, MR, Event Management	403	2%
IT-Software, Software Services	11055	41%
Recruitment, Staffing	2721	10%
Real Estate, Property	393	1%
Courier, Transportation, Freight , Warehousing	175	1%
BPO, Call Centre, ITES	1754	7%
IT-Software / Software Services	459	2%
Retail, Wholesale	294	1%
Insurance	343	1%
Pharmacy, Biotech, Clinical Research	1132	4%
FMCG, Foods, Beverage	296	1%
Education, Teaching, Training	1364	5%
Strategy, Management Consulting Firms	532	2%
Internet, Ecommerce	585	2%
Media, Entertainment, Internet	459	2%
Travel , Hotels , Restaurants , Airlines , Railways	510	2%
Automobile, Auto Anciliary, Auto Components	685	3%
Telecom, ISP	294	1%
other	424	2%
Banking, Financial Services, Broking	1778	7%
Construction, Engineering, Cement, Metals	566	2%
KPO, Research, Analytics	202	1%
Accounting, Finance	310	1%

CHART 3: INDUSTRY WISE JOB VACANCIES:



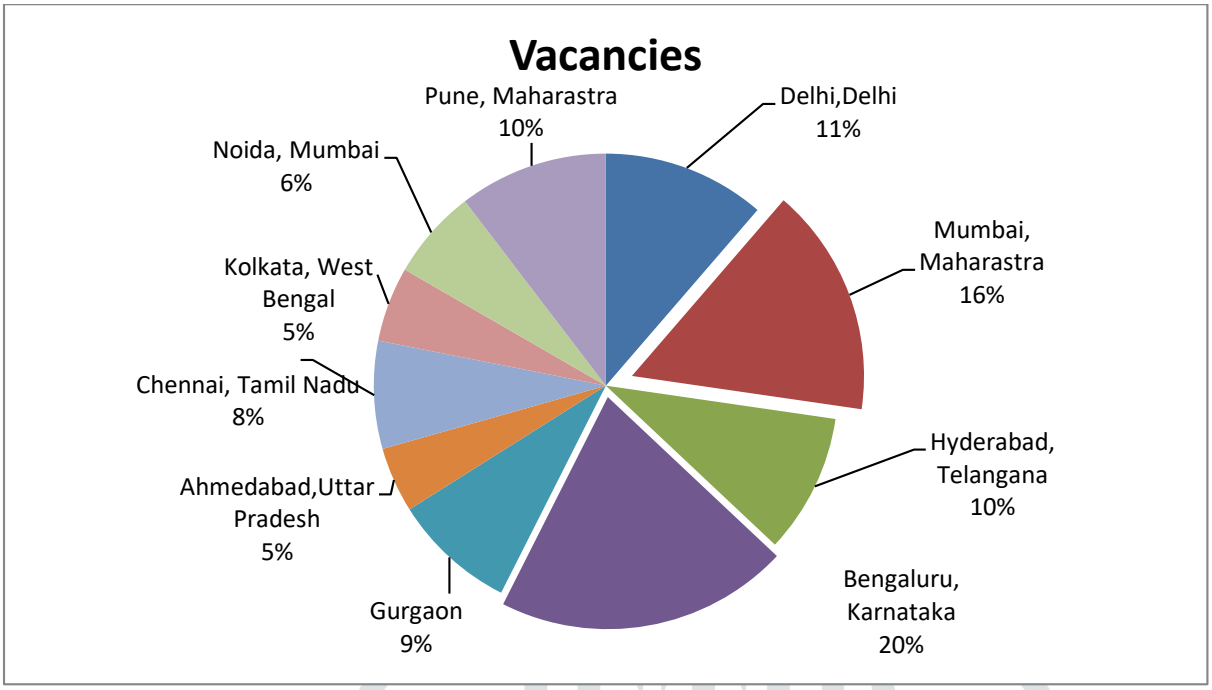
INTERPRETATION:

The chart 3 reveals that 41.35% of jobs are related to IT-Software industry only. IT –Software industry producing lot of employment opportunities compared to all other industries. And after IT industry the Recruitment, Staffing industry has 10.18% share in online job portals. After that Banking, Financial industry and Education industry providing 6.65% and 6.56% of job opportunities in online Job websites. All the other industries shown in the graph also contributing significantly to online job market.

TABLE 3: LOCATION WISE JOB VACANCIES

Location	Vacancies	Percentage
Delhi, Delhi	3803	11%
Mumbai, Maharashtra	5356	16%
Hyderabad, Telangana	3260	10%
Bengaluru, Karnataka	6859	20%
Gurgaon	2884	9%
Ahmedabad, Uttar Pradesh	1529	5%
Chennai, Tamil Nadu	2519	8%
Kolkata, West Bengal	1752	5%
Noida, Mumbai	2109	6%
Pune, Maharashtra	3484	10%

CHART 4: LOCATION WISE JOB VACANCIES:

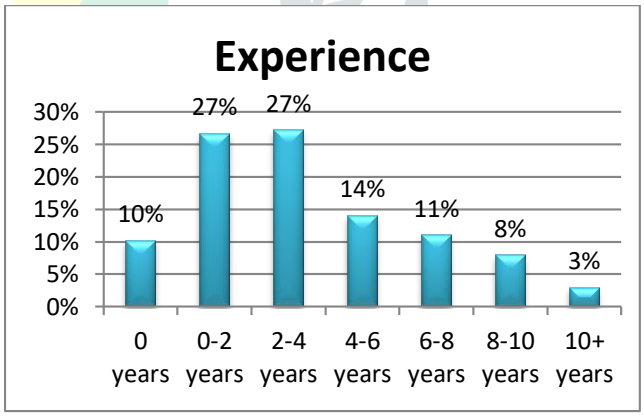


INTERPRETATION:

The chart 4 clearly shows that the available vacancies in different locations in India. In the total vacancies that are posted in online job portals Bengaluru has highest percentage(i.e. 20%) of opportunities followed by Mumbai(16%). And other locations Hyderabad(10%), Delhi(11%), Pune(10%) has offering good opportunities. Kolkata(5%), Noida(6%), Ahmadabad(5%) has offering negligible vacancies.

TABLE 4: experience wise job vacancies: CHART 5: experience wise job vacancies:

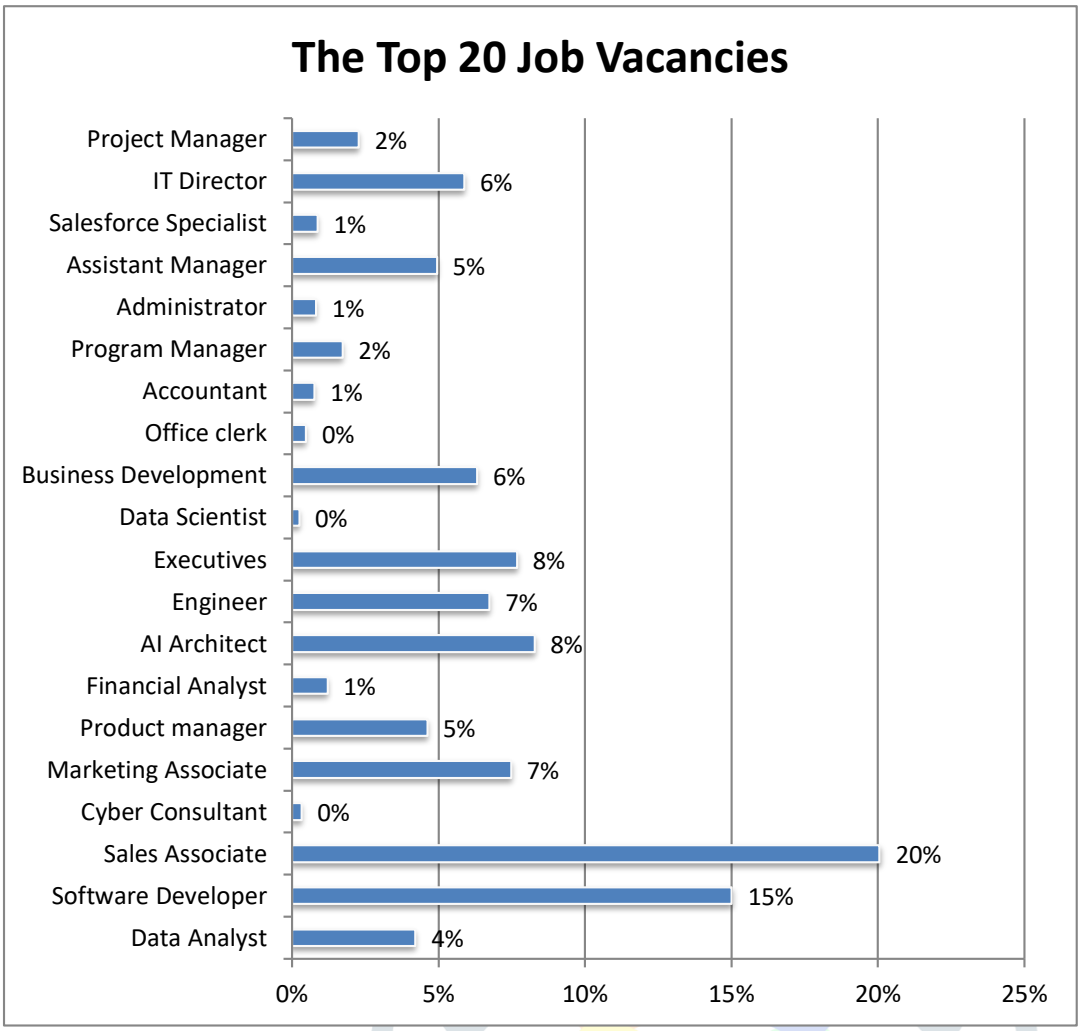
Experience	jobs	Percentage
0 years	3258	10%
0-2 years	8557	27%
2-4 years	8709	27%
4-6 years	4504	14%
6-8 years	3528	11%
8-10 years	2542	8%
10+ years	933	3%



INTERPRETATION:

The chart 5 clearly represents that the experience is one of the most main component that you need to have in order to get job. The candidates who have No experience have 10% of job opportunities in the job market, the job seekers who have at least0-2 or 2-4 years of job experience have higher demand which is 27% each in the online job market.

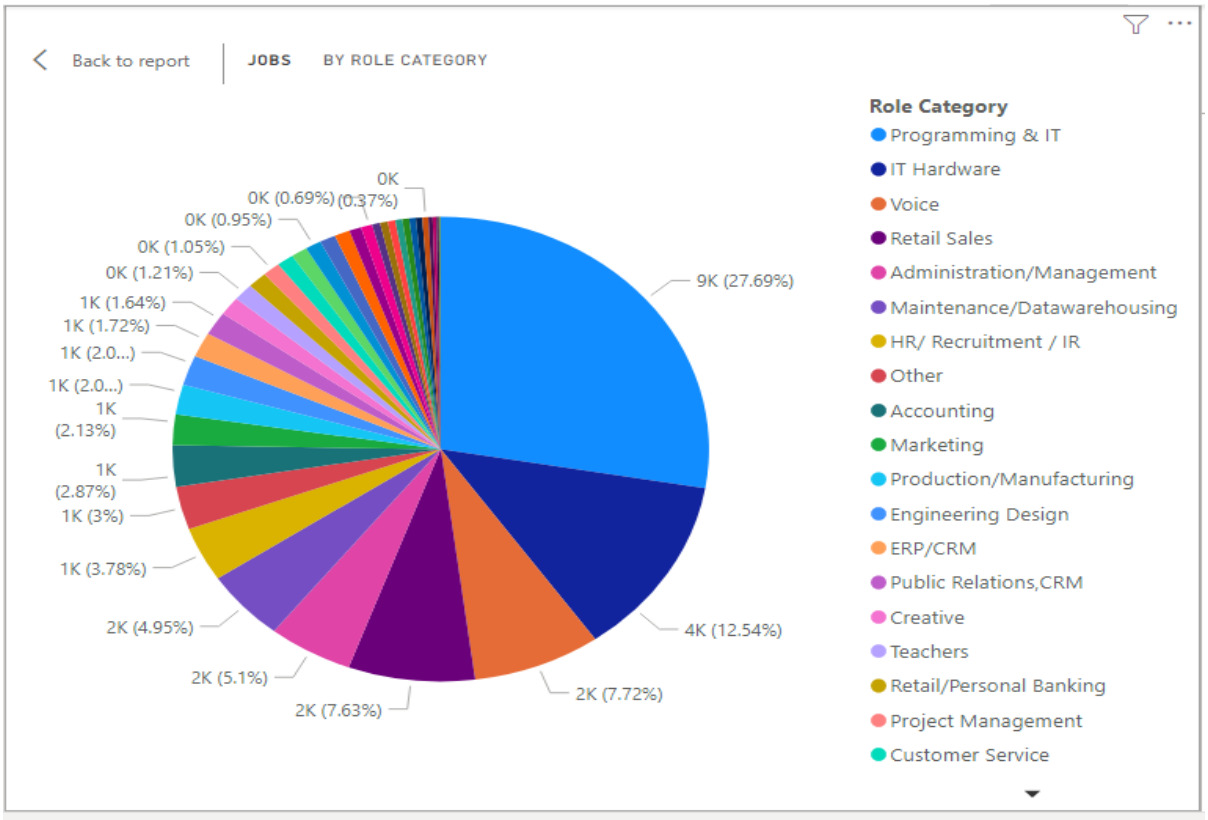
CHART 6: TOP-20 JOB TITLES:



INTERPRETATION:

The chart 6 exhibits that the best Top-20 job titles which are posted in job websites online. In those Top-20 job titles the sales associate(20%) has highest percentage of posts in the websites, and as usually Software Developer(15%) also second most featured job title in online job portals. In addition to these there many other job titles also there.

CHART 7: TOP-20 ROLES OFFERED:



INTERPRETATION:

The chart 7 reveals that the most offering roles in different companies through online job portals by recruiters are Programming & IT has highest roles(27%) among all other roles, IT Hardware role(13%), retail sales(8%), (voice) Customer Relationship(8%). And Treasury(0%), Journalists(0%) roles has offering very negligible roles to job seekers.

FINDINGS:

- It is also observed that Business Management(22%) and Sales(15%) (selling skills) has high demand among companies. It is observed that the most demanded skills are Programming languages such as Java(11%), SQL(10%),Python(4%), CSS(3%), HTML(6%), Oracle(3%), C++ and JDBC among all other skills required by Recruiting companies through online job portals.

- It is also found that IT-Software Industry (i.e. 41.35%) is providing higher job opportunities than all the other industries to job seekers. Other than IT industry the recruitment, staffing(10.06%) and Banking, Finance Industries(6.65%) are providing good opportunities to the job seekers.
- It is inferred that among many different cities all over India Bengaluru, Karnataka(20%) has most job vacancies.

- After Bengaluru the cities like Mumbai(16%), Pune(10%), Chennai(8%), Noida(6%), Gurgaon(9%), Hyderabad(10%), Kolkata(5%), and Ahmedabad(6%) offering relatively high job opportunities than other cities in India.

- It is observed that there is Top-20 Roles categories mostly offered by recruiting companies which includes Programming, IT, Sales associate, Management, HR, Accounting and Sales etc.,

- And it is found that the top-20 job titles majorly circulate in the online job websites which are majorly Software developer, Sales associate, AI architect, Data analyst, Administration / Management, and Engineers.

CONCLUSION:

The study concluded that the job seekers should learn highly demanding skills which are management, IT skills in order to get job or to attract recruiting companies. And IT-Software industry providing highest opportunities followed by Recruiting and Banking industries. Then the Bangalore is providing highest vacancies in India followed by Mumbai, Delhi being next best.

SUGGESTIONS:

- It is suggested that the job seekers should try to develop their skills according to the present recruiting companies skill requirements.
- It is suggested that the students/job seekers who have sound knowledge on programming languages, software skills has better opportunities, higher chance of getting jobs.
- The educational institutions should train students on highly demanded skills like Java, Python, SQL, Oracle, CSS, Data Management/Science etc.
- It is suggested that Bengaluru, Mumbai providing more job opportunities then the job seekers should try to catch those vacancies.
- Now a days companies are preferring online recruitment process than other print media, bill boards to reach job seekers through internet easily and cost effectively.
- It is suggested that the online job portals has increased significantly and day to day recruiters adding thousands of jobs into online job market so job seekers must go through job sites frequently to explore opportunities.

REFERENCES:

1. **S. d. S. Sirisuriya, "A comparative study on web scraping," 8th International Research Conference, KDU, p. 135–140, November 2015.**
2. **R. B. Mbah, M. Rege, and B. Misra, "Discovering job market trendswith text analytics," in 2017 International Conference on InformationTechnology (ICIT).**
3. **S. Munzert, C. Rubba, P. Meißner, and D. Nyhuis, Automated datacollection with R: A practical guide to web scraping and text mining.John Wiley & Sons, 2014.**
4. **[26] J. Ward, Instant PHP web scraping. Packt Publishing Ltd, 2013.**
5. **F. Suleman, "The employability skills of higher education graduates: insights into conceptual frameworks and methodological options," Higher Education.**
6. **A. Radermacher and G. Walia, "Gaps between industry expectationsand the abilities of graduates," in Proceeding of the 44th ACM technicalsymposium on Computer science education, 2013**
7. **L. Richardson, "Beautiful soup," Jan 2020. [Online]. Available:https://www.crummy.com/software/BeautifulSoup/**

8. S. Behnel, M. Faassen, and I. Bicking, "lxml: Xml and html withpython," 2005.

9. Source:https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/

10. https://www.google.com/url?sa=t&source=web&rct=j&url=http://www.wthtjsjs.cn/gallery/1-whjj-june

11. Source:https://www.globenewswire.com/news-release/2022/07/12/2477949/0/en/Global-Web-Scraper-Software-Market-Market-Segments-By-Type-By-Application-and-Region-Analysis-of-Market-Size-Share-Trends-for-2014-2019-and-Forecasts-to-2030.html

12. - https://www.reportlinker.com/p06191721/?utm_source=GNW

13. Wutan Huatan Jisuan Jishu Volume XVI, Issue VI, JUNE/2020 ISSN:1001-1749 Page No:2 Kaushal Parikh et. al.