Natural Language Processing for Data Science Workforce Analysis

M. Prema[1*], Dr.V. Raju[2] and M. Ramya[3]

[1*]Vice Principal, Sri Ramachandra Faculty of Engineering and Technology,
Sri Ramachandra Institute of Higher Education and Research, Chennai, India.
m.prema@sriramachandra.edu.in, ORCID: https://orcid.org/0000-0001-8970-6533

[2]Provost, Sri Ramachandra Faculty of Engineering and Technology,
Sri Ramachandra Institute of Higher Education and Research, Chennai, India.
provost@sret.edu.in, ORCID: https://orcid.org/0000-0001-8456-6615

[3]Lecturer, Sri Ramachandra Faculty of Engineering and Technology,
Sri Ramachandra Institute of Higher Education and Research, Chennai, India.
ramya@sret.edu.in, ORCID: https://orcid.org/0000-0003-3084-3490

Received: October 02, 2022; Accepted: December 17, 2022; Published: December 31, 2022

Abstract

As the demand for people with Data Science and Data Analysis skills are rising at a very high rate, periodic exploration of the skill sets for jobs in these fields have become essential. This research presents the use of Natural Language Processing for Human Resource Management. It presents the application of such techniques and tools as Python Libraries with Beautiful Soup and Selenium, Web Scrapping, Topic Analysis, Sentiment Analysis, and Natural Language Processing in the identification of skill sets related to Data Scientist, Data Analyst and Data Engineer.

Keywords: Natural Language Processing, Topic Analysis, Sentiment Analysis, Web Scrapping, Beautiful Soup, Selenium, and Data Cleaning.

1 Introduction

Technological advances in the new millennium have contributed immensely to the growth in all sectors of the global economy. Many of the emerging technologies are also expected to help address issues from elimination of poverty to improve the standard of living, from the management of water and other natural resources to reverse global warming, and from enhancing healthcare services to eradication of deadly diseases everywhere. Artificial intelligence and its applications in various industries are already changing the ways the businesses operate today. According to published reports that in the next five to seven years, AI will account for about $ 1.5 trillion of the global economy [1]. Currently growing at a rate of about 30% annually, AI is impacting the functions of every industry from healthcare to manufacturing, from automotive to telecommunication and from human resource management to supply chain operations.

AI involves a growing list of contributing technologies and scientific concepts. They include aspects of computer technologies, machine learning, robotics, internet of things, statistics, data science

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 13, number: 4 (December), pp. 225-232. DOI: *10.58346/JOWUA.2022.I4.015*

*Corresponding author: Vice Principal, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India.

and analysis, information management and more. At the core of AI is the way information is generated, managed and utilized. Data science and analysis, as important elements of AI utilize other complementing aspects of AI as machine learning, image capturing, natural language processing and others in gaining an understanding of complex systems, their operations and how they would impact the future of society. They are also at the center of broader information management process. As a result, the human resources need for professionals with skills in data science and analysis have been growing at a very fast rate around the world. Academic institutions, professional societies and governmental organizations everywhere have been doing their best to address the human resources need of data science and analysis.

This paper looks at the workforce needs of data science and analysis. In doing so, it looks at the way an important AI tool, the Natural Language Processing (NLP) can be used to extract pertinent information on skills requirement in industry and help academic institutions and industry address the challenges in workforce development.

2 Natural Language Processing in Human Resources Management

NLP is a machine learning tool which is usually tasked to look at massive data for specific texts, phrases, and repeat information leading to deciphering patterns, contexts, human behaviors, feelings, employee job performance and more. Evolving from the process of combining linguistics and computer programming, NLP is accepted as a powerful tool in industry. Such organizational data as regularly updated employee profiles, work reports, performance assessment, employee feedback, meeting minutes, complaints on working conditions, external and internal referrals, industry surveys, and professional association feedbacks are some of the sources to identify the talent, best performers, and issues on the horizon to deal with. On the other hand, such public domain information as job advertisements, social media posts, and reports from various sources may be used to determine the levels of demand for job categories, and the skill requirements under each job category [2].

When it comes to NLP applications to human resources management, there are two approaches available, namely topic analysis and sentiment analysis. Topic analysis is a machine learning technique that organizes, identifies and classifies massive data in the form of text categories by using identifiers known as “tags”. Topic analysis may use unsupervised machine learning or supervised machine learning approaches. In the unsupervised machine learning approach, known as “topic modeling” patterns in text are recognized and grouped into useful information without the application of pre-identified tags. On the other hand, in the supervised machine learning approach, known as “the text classification model”, pre-defined texts or tags are used to categorize and group text data into useful results [3, 4].

Sentiment analysis uses the employee input in the form of surveys, feedbacks, performance appraisals, etc., to measure employee’s perceptions and emotions on working conditions. Input as favorable working conditions, unfavorable working conditions or no opinion on the working conditions are used to determine if broader issues exit in an organization to deal with. Such analysis may also help identify best performing employees and their potentials for larger roles in the organization.

3 Review of Related Literature

Armin Alibasic et. al. (2022) used tools like Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), Factor Analysis and Non-Negative Matrix Factorization (NMF), to study changes in the market and identifying disparities between skills that are covered by the educational system, and the skills that are required in the job market.

Amit Verma et. al., performed pair wise comparison between online job postings related to professions such as business analyst (BA), business intelligence analyst (BIA), data analyst (DA), and data scientist (DS) using content analysis. They present a ranked list of relevant skills belonging to specific skills categories for the studied positions using Statistical and content analysis techniques [5].

Applegate inspected the data sources used for academic job ads using web scrapping technique [6]. Papou et. al., performed the mapping on knowledge extraction from online sources for Software Engineering Job Market [7].

These studies did not use identification of methods for skill identification and their granularity in different sectors. The current research focuses on identification of skill sets related to Data Scientist, Data Analyst and Data Engineer.

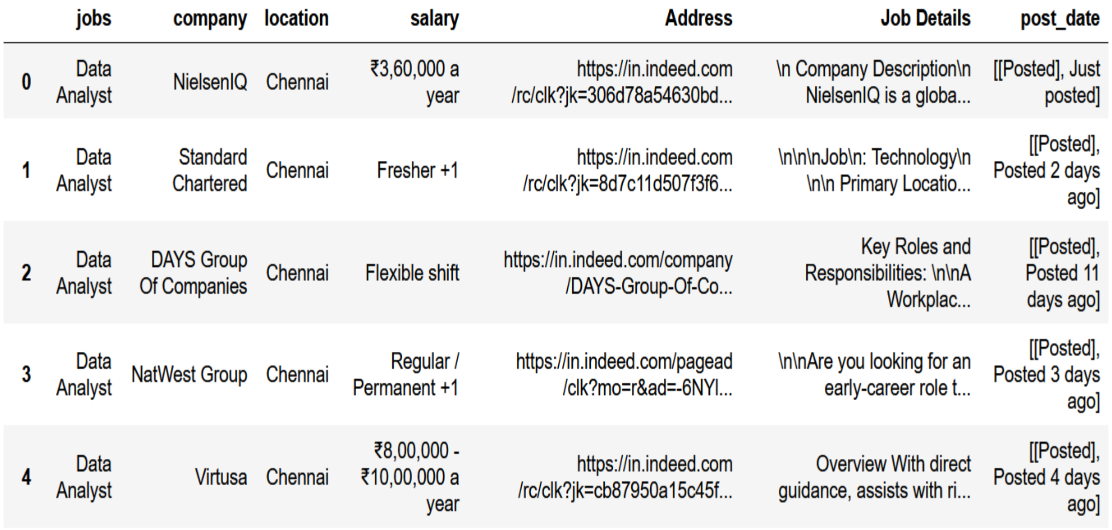
4 Research Problem and Methodology

In recognition of the growth of Data Science and Analysis as a key segment for global employment, this research was aimed at identifying the broader patterns in the demand for skills related to the area. It was also used as a means to undertake a case study approach to apply machine learning tools to human resources management. The published data from the popular public domain, Indeed.com, was used to identify the top skills needed in the industry for Data Science and Analysis. The Natural Language Processing (NLP) with Web Scrapping Technique was used as a tool for the study. Supervised machine learning approach, the text classification modeling with predefined “text tags” was used to conduct the study.

5 Application of Web Scrapping Technique

Web scrapping is defined as the process of collecting required information from web site using the python libraries like Selenium and Beautiful Soup. The collected data are converted to a data frame format after relevant data cleaning process. The process involves providing the job location and number of pages of the source data to be scrapped. The web scrapping code written using python uses Selenium driver to load the job-related web pages of such sources as Indeed, Naukri and Monster. The Beautiful Soup is used to extract such job relevant information as Job Title, Job Location, Job Description, Salary, Skill Sets, Date of Posting and other information. The extracted information is prepared into a “data frame” for further analysis.

The NLP technique is used to perform data cleaning process on the text information obtained on job description, associated skills, and other details. After data cleaning, the frequency of occurrence of skill sets associated with various job categories and job roles are extracted from the data frame. The process is extended for large data frames. A sample data frame for the Data Analyst job scrapped from Indeed.com is presented in Fig-1.



	jobs	company	location	salary	Address	Job Details	post_date
0	Data Analyst	NielsenIQ	Chennai	₹3,60,000 a year	https://in.indeed.com /rc/clk?jk=306d78a54630bd...	\n Company Description\n NielsenIQ is a globa...	[[Posted], Just posted]
1	Data Analyst	Standard Chartered	Chennai	Fresher +1	https://in.indeed.com /rc/clk?jk=8d7c11d507f3f6...	\n\n\nJob\n: Technology\n \n\n Primary Locatio...	[[Posted], Posted 2 days ago]
2	Data Analyst	DAYS Group Of Companies	Chennai	Flexible shift	https://in.indeed.com/company /DAYS-Group-Of-Co...	Key Roles and Responsibilities: \n\nA Workplac...	[[Posted], Posted 11 days ago]
3	Data Analyst	NatWest Group	Chennai	Regular / Permanent +1	https://in.indeed.com/pagead /clk?mo=r&ad=-6NYl...	\n\nAre you looking for an early-career role t...	[[Posted], Posted 3 days ago]
4	Data Analyst	Virtusa	Chennai	₹8,00,000 - ₹10,00,000 a year	https://in.indeed.com /rc/clk?jk=cb87950a15c45f...	Overview With direct guidance, assists with ri...	[[Posted], Posted 4 days ago]

Figure 1: Data Frame Derived from Web Scrapping

The figure is a snapshot of the sample data frame. It includes a list of five Data Analyst jobs, located in Chennai, India. Other details included in the data frame are company name, salary details, type of employment, the website address for the job, job roles and responsibilities and the date the job was posted.

The process is also used to derive such summary data as the companies hiring for a particular job and the number of openings in the companies in a given location.

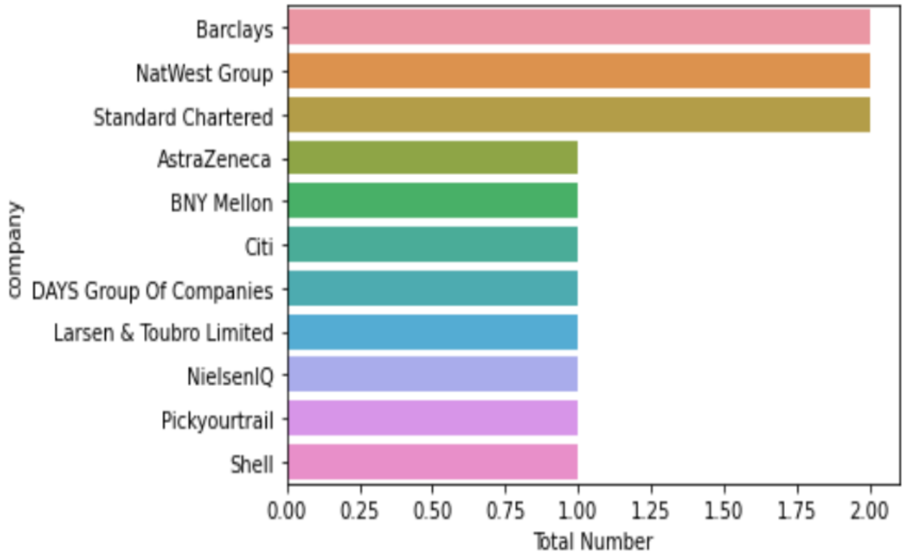


Figure 2: Data Analyst Jobs in Chennai, India as extracted from Indeed.Com

Fig-2 presents a graphical display of Data Analyst jobs posted by 11 companies located in Chennai, India.

6 Research Process and Findings

In this study, published data from Indeed.com for USA for 2018 with job roles as “Data Scientist”, “Data Analyst” and “Data Engineer” were used for data scrapping. Predefined dataset available from Kaggle was used for the purpose of “tags” and analysis.

The description of the features of the datasets are as follows;

Job Title:	Title of Role
Link:	Weblink of Job Posting
Salary:	Salary Range of the Job Posting (Estimated/Actual)
Job Type:	Categories: data_scientist, data_analyst, data_engineer
Skills:	List of desired skills
Number of Skills:	Count of the number of required skills
Days Since Posted:	Number of days since the job was posted
Description:	Web scrape of the job description
Location:	State where the job opening is listed
Company Industry:	Industry of hiring company

Table -1 presents a set of predefined skill set dictionary with skills segregated into various categories was compiled. Using this predefined dictionary, the skills for the job from Indeed.com descriptions were identified, categorized and ranked. This task was attempted with two different NLP packages: **Tex-tract and spaCy** [8].

Table 1: Pre-defined Skill Set Dictionary for Data Science and Analysis Jobs

Computer and Information Technology skills	Python','Numpy','Pandas','Matplotlib','Seaborn','Scikit-learn','PyTorch','R programming','Data structures','algorithms',
	'DBMS','Cyber security','Web programming and development','Cloud computing',
	Computer vision and Image Processing'
Statistics and Mathematics skills	Calculus','Linear Algebra','Probability','statistics','Regression','correlation', 'hypothesis','timeseries','decision making'
Project management	administration','agile','budget','cost','direction','feasibility analysis',finance','kanban','leader','leadership','management','milestones','planning',pmi','pmp','problem','project','risk', 'schedule','scrum','stakeholders'
AI & Data analytics	Data modeling','Data mining', 'Data visualization - Tableau', 'Data wrangling and preprocessing', 'Pattern Recognition','analytics','api','aws', 'big data', 'business intelligence', 'clustering', 'code', 'coding', 'Machine Learning', 'Deep Learning', 'Apache Spark', 'data', 'database', 'data mining', 'data science', 'deep learning', 'hadoop', hypothesis test', 'iot', 'internet', 'machine learning', 'modeling', 'nosql', 'nlp', 'Neural Networks', 'Keras', 'TensorFlow', predictive', 'programming', 'python', 'r','sql', 'tableau', 'text mining', 'Hadoop', 'Reinforcement Learning', 'visualuzation'
Healthcare	adverse events','care','clinic','cphq', 'ergonomics', 'healthcare',
	health care', 'health', 'hospital', 'human factors', 'medical', 'near misses',
	patient', 'reporting system', 'Drug Development processes', 'Patient care operations', 'Healthcare regulations', 'Human resource functions in healthcare facilities
Organization and Leadership Skills	Interpersonal communication', 'Team work', 'Decision making', 'Lifelong learning', 'Ethical practice

The research resulted in skill sets as “data frame” in multiple pages. The text from each page was extracted using “Tex-tract". Data cleaning was done by removing white spaces, special characters, numbers and punctuation. Each word from the extracted text were matched with the skillet dictionary to arrive at the final results.

7 Data Science and Analysis Skills: NLP Application Results

The results from the application of NLP to Indeed.com jobs posting in the USA for the year 2018 for “Data Scientist”, “Data Analyst” and “Data Engineer” are presented in Table -2. It is seen from the data that the AI and Data Analytics Skills, Organization and Leadership Skills, and Project Management Skills were considered to be the most prominent with more than 20% to 30% of jobs requiring these skill sets.

Table 2: Results of NLP Application

Skill Category	**Percentage of Jobs with the Skill Category**
AI and Data Analytics Skills	29%
Organization and Leadership Skills	29%
Project Management Skills	12%
Healthcare Skills	12%
Statistics Skills	6%
IT Skills	3%

With the help of **spaCy,** a library for advanced Natural Language Processing in Python, Rule based matching was performed to identify the top skills required for a data scientist. By matching the skills using the skill set_dictionary_corpus to the words extracted from description column, top skills for the role of data scientist was identified along with their frequency and score. The results are presented in Table -3

Table 3: Top Skills for the Role of Data Scientist

Skill	**Count**	**Percentage**
data	5236	0.9161854768
python	3238	0.5665791776
sql	3035	0.5310586177
analytics	2857	0.4999125109
management	2587	0.4526684164
r	2163	0.3784776903
statistics	2111	0.3693788276
machine learning	2111	0.3693788276
modeling	1870	0.3272090989
programming	1843	0.3224846894
data science	1635	0.2860892388
hadoop	1570	0.2747156605
problem	1526	0.2670166229
big data	1492	0.2610673666

From the result of rule based matching, the key skills required are determined as data handling, python, SQL, analytics, management, R and statistics, machine learning, modeling, programming, hadoop, problem solving, and big data handling. It is quite evident from the analysis of the job postings in Indeed.com that the potential job seekers for jobs as Data Scientists, Data Analyst or Data Engineer should have the following skill sets: programming languages with R and Python, Data Base Management, Data Visualization Techniques, Proficiency in Statistical Analysis, Machine Learning Skills, Use of Python libraries and an overall understanding of the role of data in management.

The contribution of this research is to demonstrate the application of Natural Language Processing to workforce skills analysis. It is hoped that this work would help others to look at the application of the techniques for further study of skills needed in various fields. For the purposes of further study and explorations, the authors may be contacted for the coding developed and used in this study.

8 Conclusions and Suggestions for Further Study

This research was limited to the application of NLP to identify skills required for jobs as Data Scientists, Data Analyst and Data Engineer. It was also limited to the application of NLP to data available in the public domain. Repeating the study for multiple years, and utilizing data sets from a number of sources would further validate the results. It is also possible to undertake an expanded study utilizing other AI and machine learning tools to project the skills requirements.

Acknowledgment

The authors would like to thank Prof. P. Angayarkanni for her valuable suggestions and input towards workforce analysis.

References

[1] Fortune Business Insights, “AI Market Size by 2029” Report ID: FBI100114, Artificial Intelligence Market Growth, Trends, Forecast, 2029 (Fortunebusineeinsights.com).

[2] Sengupta, R. (2022). How Natural Language Processing can Revolutionize Human Resources, Academy to Innovate HR.

[3] Lexalytics Whitepaper, (2022). Machine Learning for Natural Language Processing and Text Analytics. https://www.lexalytics.com/resources/ml-nlp-whitepaper/

[4] Guenole, N., & Feinzig, S. (2018). The business case for AI in HR. *With Insights and Tips on Getting Started. Armonk: IBM Smarter Workforce Institute, IBM Corporation*. https://www.ibm.com/downloads/cas/AGKXJX6M

[5] Alibasic, A., Upadhyay, H., Simsekler, M.C.E., Kurfess, T., Woon, W.L., & Omar, M.A. (2022). Evaluation of the trends in jobs and skill-sets using data analytics: a case study. *Journal of Big Data*, *9*(1), 32.https://doi.org/10.1186/s40537-022-00576-5

[6] Applegate, R. (2010). Job ads, jobs, and researchers: Searching for valid sources. *Library & Information Science Research*, *32*(2), 163-170. http://www.sciencedirect.com/science/article/pii/S0740818810000071

[7] Papoutsoglou, M., Ampatzoglou, A., Mittas, N., & Angelis, L. (2019). Extracting knowledge from on-line sources for software engineering labor market: A mapping study. *IEEE Access*, *7*, 157595-157613. https://ieeexplore. ieee.org/abstract/document/8884193

[8] Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science*, *5*(1), 861-882. https://doi.org/10.1007/s42001-021-00149-1

[9] Choudhary, A., Choudhary, G., Pareek, K., Kunndra, C., Luthra, J., & Dragoni, N. (2022). Emerging Cyber Security Challenges after COVID Pandemic: A Survey. *Journal of Internet Services and Information Security, 12*(2), 21-50.

Authors Biography

M. Prema is a faculty member and administrator in the Faculty of Engineering and Technology at Sri Ramachandra Institute of Higher Education and Research, Chennai. Prema is the current Vice Principal of Engineering and Technology at Sri Ramachandra Institute of Higher Education and Research. In the last twenty years, she has been involved in the development and implementation of engineering and technology degree programs in India and the USA, implementation of innovative approaches to teaching and learning, and helping students prepare for careers in the broader global industry. Her research interests are in technological skill development in industry, and the application of statistics, machine learning and data analytics to study complex problems.
Email ID: m.prema@sriramachandra.edu.in
ORCID: https://orcid.org/0000-0001-8970-6533

Dr.V. Raju is a faculty member and senior administrator in the Faculty of Engineering and Technology at Sri Ramachandra Institute of Higher Education and Research, Chennai. Prof. Raju is the Provost at Sri Ramachandra Institute of Higher Education and Research. He was responsible for the establishment of the Faculty of Engineering at SRIHER and creating bachelors, masters and doctoral programs in Computer Science and Engineering with specializations in Artificial Intelligence, Machine Learning, Data Science and Analytics, Cybersecurity, Internet of Things, Medical Engineering, and Medical Bioinformatics. His current interests are in the studies relating to advanced technologies, experiential education, and the impact of technology, and policy on workforce development.
Email ID: provost@sret.edu.in
ORCID: https://orcid.org/0000-0001-8456-6615

M. Ramya is a faculty member in the Faculty of Engineering and Technology at Sri Ramachandra Institute of Higher Education and Research, Chennai. She began her career in industry before turning to academia and teaching undergraduate and graduate courses in Artificial Intelligence, Machine Learning and Data Analytics. Her current research interests are in the application of Machine Learning tools and techniques to a wide range of problems, including healthcare studies.
Email ID: ramya@sret.edu.in
ORCID: https://orcid.org/0000-0003-3084-3490