An empirical study on Natural Language Processing in identifying industry demands to recommend required qualifications in the UAE

دراسة تجريبية حول معالجة اللغة الطبيعية في تحديد متطلبات الصناعة للتوصية
بالمؤهلات المطلوبة في دولة الإمارات العربية المتحدة

by

KANNAN SUSEELAN UNNITHAN

Dissertation submitted in fulfillment
of the requirement for the degree of
MSc INFORMATICS
(Knowledge & Data Management)
at
The British University in Dubai

May 2021

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above, for which I am the author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

Abstract

A skillful and competent workforce is one of the deciding factors of any countries economic and social growth. Developed and undeveloped countries have issues related to the availability of skilled laborers. Also, the existing workforce is struggling to upgrade their skills based on the fast-changing technologies and industry environments. These factors are boosting the unemployment rate of countries. The solution for this issue is Vocational Education and Training, which will equip the workforce with the right skills. The vocational qualifications are developed based on the industry requirements. Regularly identifying and updating the industry requirements is a manual process and is expensive and time-consuming. By extracting and processing the employment advertisements published on the web, one can analyze the skills demands in each industry and the organization's operational effectiveness. This research oversees the application of web data extraction in the education sector. This study proves that by extracting the jobs published on the web and processing the same using the NLP technique, the industry's skill requirements can be identified and propose new qualifications to be developed.

خلاصة

القوة العاملة الماهرة والمختصة هي أحد العوامل الحاسمة في النمو الاقتصادي والاجتماعي لأي دولة. البلدان المتقدمة والمتخلفة لديها قضايا تتعلق بتوفر العمالة الماهرة. أيضًا ، تكافح القوى العاملة الحالية لترقية مهاراتها بناءً على التقنيات سريعة التغير وبيئات الصناعة. تعمل هذه العوامل على زيادة معدل البطالة في البلدان. الحل لهذه المشكلة هو التعليم والتدريب المهني ، الذي سيزود القوى العاملة بالمهارات المناسبة. يتم تطوير المؤهلات المهنية بناءً على متطلبات الصناعة. يعد تحديد متطلبات الصناعة وتحديثها بشكل منتظم عملية يدوية ومكلفة وتستغرق وقتًا طويلاً. من خلال استخراج ومعالجة إعلانات التوظيف المنشورة على الويب ، يمكن للمرء تحليل متطلبات المهارات في كل صناعة والفعالية التشغيلية للمؤسسة. يشرف هذا البحث على تطبيق استخراج بيانات الويب في قطاع التعليم. تثبت هذه الدراسة أنه من خلال استخراج الوظائف المنشورة على الويب ومعالجتها باستخدام تقنية البرمجة اللغوية العصبية ، يمكن تحديد متطلبات مهارة الصناعة واقتراح مؤهلات جديدة ليتم تطويرها.

Table of Content

List of Figures

List of Tables

1. Introduction

This section outlines the overview of the research. This introduction covers the research problem, discussion on the motivation, a clear research objective, research questions, and methodology. Also, a summary of the various sections of this paper is available provided here.

1.1 Overview

Any facts or known things can be referred to as data, and it is available everywhere. In this era of technology, every industry is depending on data for decision-making and process improvement. Data could be available in the form of structured or unstructured format. These days, the internet is the primary source of data for any industry, and the majority of data from the internet is in text format, which is unstructured. Natural Language Processing (NLP) helps to extract information from the text data.

Analyzing web data is challenging, and the NLP provides different methods and techniques to extract and analyze web data. Web data extraction is a hot research topic in almost every industry. Many types of research are currently going on regarding the different approaches and applications of web data extraction in various sectors.

The economy and employment rate of any country are directly related to the competent workforce available in the industries. Vocational education produces ready-to-work candidates with industry-specific skills and knowledge. Since no further training or investments, required organizations always prefer vocational education graduates. However, in UAE, technical education is relatively new, and the employers' expectations are not satisfied. Also, there is a considerable gap between the market demand and available workforce (Rahman & Al 2016). The vocational education sector is getting promoted in UAE due to the high demand and shortage of skilled workers. An effective and efficient vocational education system can produce a competent and ready-to-work workforce to support the UAE's industry demand (Salem & Alawi, 2016).

In the qualification development and quality assurance processes capturing industry requirements is the main element. The impact of qualification is measured based on how well it satisfies the industry requirements. National Qualification Authority (NQA) is the regulatory authority in UAE that manages the vocational education sector and qualification development. The NQA continuously monitors the industry and liaise with various industry representatives to identify the workforce demand. Based on the industry demands, the NQA decides on the qualification development. Identifying the need for the qualification is the first step in developing a qualification. The NQA continuously monitors industry requirements and skill demands to provide suitable qualifications to equip the workforce with the required skillsets. Continuously monitoring and identifying the industry requirements is a manual process and is expensive and time-consuming.

Jobs posted on the online job portals are a significant data source for determining the skills requirements in various industries. Performance of a company, the required skill demands in any specific sector, employee skill development requirements, etc.. can be identified by extracting and analyzing the data from the job portals.

This research identifies the advantages of the NLP in the education sector by focusing on the applications of web extraction techniques in qualification development. The study recognizes the importance of web extraction by surveying various literature. Also, this paper proposes a semi-automated method to identify the required qualifications to be developed in the UAE Information

Technology industry sector by extracting UAE IT job requirements from the online employment portals.

1.2 Problem Identification

The main challenge in any country is to have enough skilled and reliable workforce for the industries. The competent workforce in various industries will make any country internationally competent. To support the National interests and improve productivity, the governments need to effectively implement multiple strategies to produce efficient taskforces suitable for the fast-changing industry environments. One of the critical factors in maintaining the status of 'Internationally Competent' is the availability of relevant qualifications that empower the citizens with skills that tie with the highly evolving system, resources, workplace environments, and technologies. These practical tactics will significantly improve the workforce quality and productivity. And will enhance the country's well-being by producing excellent cost management and workforce management systems (NQA 2012).

Puckett, Davidson & Lee (2012) state that most countries, including the developed ones, face issues related to the scarcity of skilled laborers pertaining to the specialized and technical sectors causing high unemployment rates. The technical and specialized positions in the industries are not able to fill quickly due to the lack of a workforce with the required skills and knowledge training. Even the potential staff are not able to manage the specialized tasks due to the unavailability of effective and efficient training, which will cause job loss of existing employees and increase the unemployment rate. The solution for these issues is Vocational Education and Training. If implemented and managed correctly, vocational education can significantly reduce the gaps between the industry demand and workforce supply and reduce the unemployment rate.

Esposito, ElSholkamy & Fischbach (2017) pointed out that the future of UAE industries is highly dependent on a heavily skilled workforce. They also mentioned that to cope with the quickly evolving systems, technologies, and to satisfy the fast-changing industry requirements, proper strategies have to be implemented by the UAE. These strategies should effectively support the skills development of its citizens and improve productivity. Generally, employers prefer vocational education graduates as they are well prepared with ready-to-work skills, and no further training is required. However, in UAE, technical education is relatively new and under the development stages. In terms of the available workforce. The employers' expectations are not satisfied. There is a considerable gap between the industry requirements and the skills obtained by the graduates (Rahman & Al 2016). To manage the scarcity of skilled employees, the UAE government is bound to provide high priority in encouraging vocational education among the nationals.

According to Salem & Alawi (2016), by establishing an effective vocational education sector, a reliable and competent workforce could be created to support the requirements of the national and international organizations in the UAE, which will improve economic enhancement by increasing productivity. Due to the high demand and shortage of a skilled workforce, the vocational education sector is getting promoted in UAE. Rahman & Al (2016) points out the major problems that the vocational education sector faces in UAE. These primary challenges are Education quality, Acceptance of vocational education, Globalization and Growing demand, Graduates and market alignment, and Lack of qualified staff. This study picks the Graduates and market alignment from these five challenges to identify and propose a solution.

Ensuring the effectiveness of vocational education is a cyclic process, and it starts from the identification, planning, and development of qualifications. Quality of qualification is referred to its usability, consistency in the outcomes, return on investments, stakeholder satisfaction, and continuous improvement. A qualification’s quality is identified by its contribution to the learners’ skill development, increased employability, and progress in innovations and productivity (KHDA 2018). Four aspects of qualification quality assurance are Inputs, Processes, Outcomes, and Impact. The factor ‘Input’ deals with the factors that identified the necessities of the qualification, such as government policies and strategy, financial sustainability, provisions, and industry requirements. The second factor, ‘Processes’ deals with the operational features such as candidate selection, qualification delivery, assessments, internal and external quality assurance, and student satisfaction. The third factor, ‘Outcomes’ handles the students’ success rate in terms of grade, and progression & continuous improvement. Finally, the fourth factor, ‘Impact’ deals with learners' employability, career progression, and industry satisfaction.

Capturing industry requirements is the main element in the qualification development and quality assurance processes. The success of a qualification relies on how well the industry requirements are captured and the extent to which the qualification satisfies the industry demands. In UAE, the National Qualification Authority (NQA) is responsible for developing and endorsing nationally recognized vocational qualifications. The qualification development includes a high standard and integrity process. The qualifications are developed based on the current requirements and demands shown in each industry sector. The NQA liaise with various industry sectors to capture their demands and conditions related to the workforce. According to the identified needs, NQA works closely with the industry team to develop multiple vocational qualifications.

The first step in developing a qualification is the process of identifying the need for the qualification. To provide the required staffing on time for each industry, the authorities must continuously monitor the industry requirements and skill demands. It takes six months to 1 year or more to develop and implement vocational qualifications. The main bottleneck in the qualification development process is the identification of industry requirements. One of the main risks in the qualification development is that, due to the continued advancement in technologies and industrial processes, the needs might have been changed by the time the relevant bodies approve the qualification for delivery.

Regularly identifying and updating the industry requirements is a manual process and is expensive and time-consuming. This research examines the advantages of applying Web Scrapping methods to determine the current industry demands and propose required qualifications by extracting the job requirements posted on the online job portals in UAE.

1.3 Inspirations

According to Lloyd (2008), the current digital world is growing to the next decade, at a rate of 40% of increase every year. Individuals and organizations, the internet inhabitants, are included in the growth. This phenomenon resulted in extraordinary growth in online data and opened a new source of opportunity for information gathering. One of the top research areas in Natural Language Processing is information extraction from online sources.

Many kinds of research have been conducted in various industries, like hospitality (Han & Anderson 2021), food processing (Hillen 2019), finance (Krotov & Tennyson 2018), the stock market (Maurya et al. 2019), OGC web services (Li, Wang & Bhatia 2016), retail (Jorge et al. 2020), architectural engineering (Yang et al. 2020) (Hong, Lee & Yu 2019), textiles (Menke & Giehl 2021) and agriculture (Nashipudi 2020) to identify the application of web data extraction using the NLP techniques to improve the operational efficiency. Educational organizations can improve efficiency by implementing various NLP techniques (Alhawiti 2014).

According to Gupta et al. (2016), web scraping is an instrument to extract the required information from web resources. Using the web scraping technique, unstructured data from the web, in the form of HTML or XML, can be extracted and convert the same to structured formats like CSV or database. Phaphuangwittayakul et al. (2018) conducted a study on analyzing the labor market's skills requirements using web scrapping. In their research, Phaphuangwittayakul et al. (2018) extracted the job description from various job portals to identify the skills and qualifications required for performing the listed job title. They used the extracted data to determine the skill mismatch.

In 2020 Pillai & Amin conducted a study to identify the skills required for various IT jobs. They used web scrapping to extract skills information from the jobs posted in India's various IT job portals. In the conclusion of their paper, Pillai & Amin (2020) state that educational institutions can improve their curriculum by regularly identifying the required skills. However, their study doesn't provide any direct link or comparison between the identified skills and the existing curriculum.

The National Qualification Authority develops National qualifications in conjunction with the UAE's industry experts and representatives (VETAC Q + NOSS System Guidelines - VETAC's Quality Assurance and Endorsement System to develop and deliver National Qualifications 2014). The first step in creating a National Qualification is identifying the industry requirements and deciding on the type of qualification to be developed. The industry requirements are determined by surveying and interviewing the industry representatives. The specified requirements will then be mapped with the International Standards Classification of Occupations (ISCO) and based on the mapped ISCO, the type and level of qualification will be decided. The ISCO, is the most accepted occupation category system used Worldwide. The ISCO framework establishes the duties to be performed by a person in a particular job (Lovaglio et al. 2018).

The manual process of identifying the industry requirements and qualification identification is very lengthy and costly. The motto of this research is to examine the application of web scrapping in recognizing the UAE industry requirements to identifying industry-specific required qualifications.

Inspired by Gupta et al. (2016) and Pillai & Amin (2020), this research is intended to identify the advantages of web scraping in proposing new qualifications for the UAE industries. The literature review shows that the industry requirements can be identified by extracting and analyzing the jobs posted on the online job portals. There are many studies dealing with the approaches and applications of job information extraction from the job portals and employee web sites. However, the application of online job data in the process of qualification development was not covered much. Most of the studies identified the skills required for perfoming industry tasks by extracting and analyzing the online employment data. But the using these identified skills to proposing new vocational qualification was not covered in the available literatures. This paper will try to identify the required vocational qualifications by extracting advertised job details from the online job portals.

1.4 Research Objectives

Every industry advances as fast as technology develops. These dynamics in industries demand a skillful and up-to-date workforce. The requirements of each sector change every day. From layman to higher management need to have current knowledge about the industry. So the main challenge in any industry is to obtain suitable and expert employees. Vocational education is expected to equip the workforce with the required skills.

Vocational education prepares people for the technical sector and makes them work as technicians in various industries. The vocational qualifications are developed based on the industry requirements. Six months to one year is the normal time frame to develop and implement vocational qualifications. Identification of industry requirements is a critical factor in the qualification development process. Due to the continued advancement in technologies and industrial processes, the needs might have been changed by the time the relevant bodies approve the qualification for delivery.

Capturing industry requirements is one of the primary and most time-consuming elements in the process of qualification development. Currently, the industry demands are captured manually. The qualification development team collects the industry requirements by conducting surveys and face-to-face meetings with the industry representatives. However, the chances of missing any relevant stakeholder are very high in this manual process and can result in capturing biased industry requirements.

Many studies proved that the jobs posted on the job portals represent the actual demand of industries and can rely on identifying the industry requirements. Based on the studies conducted by Dadzie et al. (2018), Johnson, Albizri & Jain (2020), and Pillai & Amin (2020), this paper tries to capture the industry demands for qualification development by extracting job requirements posted on the web.

The main objectives of this study are:

1. Propose a semi-automated technique to propose new vocational qualifications
2. Identify the impact of web extraction techniques in gathering the industry requirements for developing UAE national qualifications
3. Identify the application of web data extraction in various industry sectors

1.5 Research Paper

This paper consists of 8 sections. Section 1 provides an overview of the research, with the background, problem identification, research inspiration, objectives of the research, and a paper structure summary. A survey of previous works is covered in Section 2. The previous works are categorized into three areas, literature review on web data extraction, application of web extraction in various industries, and job demand extraction.

The research methodology is covered in Section 3 with the details of Data Collection. Results and Discussions are covered in section 4. Section 5 provides the conclusion of the study. The limitation and future work related to this paper are covered in section 6. Appendix and References are covered in sections 7 & 8, respectively.

2. Previous works

A detailed survey of various literature related to web data extraction has been done for this research. The literature survey covered the general aspects of web data extraction, the application of web data extraction in various industries, and the advantages & application of online job information posted on the job portals.

2.1 Background

Data is a mandatory element for any research to validate the findings of the study. However, 79% of the available data is in amorphous form. Every second, millions of online data have been generated from tweets, social networks, personal messaging apps, blogs, and various industry websites. The majority of these online data are in unstructured text formats (Bansal 2017). Using Natural Language Processing and the supporting components, one can process that unstructured text to extract any required information.

In English, verb, adverbs, nouns, pronouns, conjunctions, interjections, adjectives, and prepositions are the parts of speech, and any NLP technique must understand the language structure. The main components of language that need to know by an NLP program are: Phonological, which is the sound’s most minor unit, and it relates the words to sounds; Morphological is the meaning’s smallest unit, and it is about generating words from morphemes; Syntactic is the understanding of developing accurate and sensible sentences by correctly organizing the words; Pragmatics deals with using sentences in different contexts, and the sentence meaning in according to the contexts; and World which managed the conversation by understanding the sentences (Surabhi 2013).

Morphological analysis, Syntactic analysis, Semantic analysis, Discourse analysis, and Pragmatic analysis are the steps involved in natural language processing. The first step in Morphological analysis is to identify the sentences and words. This task is referred to as Tokenization, and after this step, lemmatization is done to remove the affixes. Next, for the Syntactic analysis, the Part of Speech (POS) tagging and Syntax checking will be done in the identified sentences to check the grammar rules. Using the Word sense disambiguation classification, the Semantic analysis will be carried out to determine the words meaning based on the context of the sentences. For situations where the Semantic analysis is not able to provide accurate meaning, Discourse analysis will be implemented using Reference Resolution to interpret the correct definition. Finally, a Sacrcasam Detection task will be conducted during the Pragmatic Analysis to detect any touches of sarcasm used in the sentences (Tatwadarshipn 2021).

Initially, the Natural Language Processing researches were based on the rules of linguistics and language patterns. With the complexity of managing the new entries in the dictionaries and the complex grammar rules, the researchers started to rely on corpus-based NLP researches (Isahara 2007). A corpus is a collection of large structured, machine-readable texts formed from single or multiple languages. The usefulness of any corpus depends on its design. A perfect corpus design will be of finite size with a sample of langue text that comprises a widespread of text types. Also, the design should ensure that the corpus includes the sample that represents the populations' variability (*NLP - Linguistic Resources* 2021).

Python is a popular high-level programming language that can be used for Natural Language Processing. The main library used in Python to process and extract information from human language is NLTK. NLTK is a free, open-source library compatible with Mac, Windows, and Linux operating systems. NTLK has a very user-friendly interface to manage various corpora with libraries to support the text processing operations like parsing, classification, tagging, tokenization, and stemming. It also supports the reasoning of semantics and context-based discussion of conversations. Operations like Part-of-speech tagging, Entity Extraction, Tokenization, Stemming, Parsing, Semantic Reasoning, and text classifications can be carried out using the NLTK libraries.

A few of the other Python libraries used for the analysis of natural language are GenSim, SpaCy, CoreNLP, TextBlob, AllenNLP, Polyglot, and Scikit-Learn (Akshay 2021). GenSim is a library that uses the topic modeling and vector space modeling tool kits to identify any two documents' semantic similarity. GenSim uses algorithms that are memory-independent regardless of the size of the corpus. GenSim will support high-size inputs that can be larger than the RAM size. Due to its capability of managing the bigger size of inputs, GenSim provides high processing speed and excellent memory usage optimization. Algorithms used in GenSim for the natural language processing are Random Projections(RP), Hierarchical Dirichlet Process(HDP), Latent Dirichlet Allocation(LDA), and word2vec deep learning or Latent Semantic Analysis(LSA/SVD/LSI).

The open-source library SpaCy manages a very massive amount of text data and is implemented for building real-world projects by using production usage. To handle the enormous data, this library was developed using Python in Cython. The main advantages of SpaCy library are: faster than other libraries; it supports multiple trained transformers; supports tokenization for over 49 languages; supports segmentation of text, classification of text, tagging of part-of-speech, named entity recognition, and lemmatization; and supports 17 languages with over 55 trained pipelines.

The Stanford university provided the library CoreNLP. This library supports text processing mechanisms like part-of-speech, sentiment analysis, parser, named entity recognition, bootstrapped pattern learning, coreference resolution, lemmatization..etc., using very few lines of codes. This library also supports multiple languages like English. Arabic, Chinees, french and Spanish. Though this library was developed in Java, it supports integration with Python. TextBlob is also an open-source python library that is NLTK powered. This library gives an interface to easily process the operations like parsing, sentimental analysis, phrase and word frequencies, Part of speech tagging, spelling correction, N-grams, classifications, Tokenization, extraction of the noun phrase, and integration of WordNet.

One of the top libraries used in Python for Natural Language Processing is AllenNLP which easily handles both complex and straightforward operations. For processing the data, the open-source library SpaCy is used in AllenNLP. This also deals with lay cycles. Event2Mind is the model used in AllenNLP, which makes the library more suitable for client purpose response and service advancement. Polyglot is an NLP library that is not popular as other libraries. However, Polyglot is a powerful library that supports the analysis of more languages. This can manage many languages that other libraries can't. Polyglot supports 165 languages for Tokenization, 196 languages for Language detection, 40 languages for Named Entity Recognition, 16 languages for Part of speech tagging, 136 languages for sentimental analysis, 137 languages for word embedding, 135 languages for Morphological analysis, and 69 languages for Translation.

Another open-source library used in Python is Scikit-Learn. This library is famous among data scientist as it includes a variety of algorithms that supports machine learning models. The models used in this library have great intuitive methods and provides a wide range of documentation to support the data scientists. Scikit-Learn supports the conversion of text into numerical vectors using various bag-of-words functions. However, the neural network for text processing is not used in this library and is not fit for carrying out complex tasks like part of speech tagging for text corpora.

The process of extracting information from the data collected from the web is referred to as Web Mining. Three categories of web mining are web usage mining, web content mining, and web structure mining. Web user's pattern and behavior is captured in web usage mining. The connection between various web pages and their links are identified using web structure mining, and the web content is collected using web content mining (Selvadurai 2013).   The first step in web data mining is identifying the information to be extracted from the web and providing input accordingly. Then the system will analyze the input and find the related information nodes to be extracted (Chen 2010).

A programmed strategy to extract data from the web is generally termed as Web Scraping. Required information from a specific website or general information from various websites can be captured using web scrapers. The majority of the data present on the web is in an unstructured format. These data can be extracted, organized, and then stored in the desired form using web scrapping. Many methods are available to extract data from the web. These methods include Application Program Interface (API), web administrations, manual software coding, and ready-made applications. Many of the popular web application providers like Google, LinkedIn, Facebook, Twiter, StackOverflow, etc., provide APIs to extract information from their web pages.

By using web scraping, all information or any specific information from a web page can be extracted. The first step in web scraping is to specify the URL of the target web page, and the scraper will load all the HTML sources of the target web page and may process the Javascript and CSS part of the HTML. Then the web scraper method will take all necessary information from the required HTML tags or CSS and convert that information into a structured format like JSON, CSV, or excel.

Python provides multiple libraries like Requests, Beautiful Soup 4, lxml, Selenium, and Scrapy to implement web scraping quickly (Akshay 2021). Also, Python has a dedicated package called

Natural Language Tool Kit (NLTK) to support Natural Language Processing. NLTK processes strings by accepting strings as input, and the output will be a string or an array of strings (Yash 2020).

The main methods in web scraping are Mimicry Approach, Weight Measurement Approach, Differential Approach, and Machine Learning Approach. Using these methods, many tools have been developed to extract data from the web automatically. Few of such ready-made web data extraction tools are, Spider, Parse Hub, Data Scraper, Agenty, Data Miner, Cloump U-Scraper Plugin, OutWit Hub, and Dexi.IO (Diouf et al. 2019)

Different methods of web scarping are manual data extraction using simple copy and paste, custom coding, and web scraping tools like ParseHub. ParseHub is a free and powerful web scraper tool to extract employment data from the job portal. The advantage of using ParseHub is that it is effortless to extract data from any website, especially for data spread over multiple pages, and have complex structures implemented using javascript or CSS. The main features of ParseHub are:

- Scheduled Run: ParseHub support the option to schedule the web scraping in regular intervals or for a specific time.
- Automatic IP Rotation: If the website has a web scraper blocker, then the parse hub will automatically implement the IP rotation to overcome any web scraping blocker.
- Ajax and Javascript Enabled interactive websites: Using ParseHub, it is easy to extract data from interactive websites developed using Ajax or Javascript
- Expressions and Conditionals: ParseHub supports the Expressions and Conditionals used on the web pages.
- CSS, XPATH, RegEx Selectors: Using ParseHub, it is easy to extract web pages developed using CSS, XPATH, RegEx
- Tables and maps: ParseHub can easily extract data from tables and maps
- Extract HTML, text, and attributes: It is very easy and straightforward to extract HTML data

- Download images/files: ParseHub supports the scraping of images and various files types
- Access Secured web data: Using ParseHub, data from login secured web pages can also be accessed.
- Scrolling pages: ParseHub supports the extraction of data from unlimited scrolling pages.
- Forms and inputs: Using ParseHub, data available on the HTML forms or input fields can be extracted easily.
- Dropdowns, tabs, and pop-ups: Normally, accessing data from Dropdowns, tabs, and pop-ups are very difficult. But by using ParseHub, data from Dropdowns, tabs, and pop-ups can be easily extracted.
- Paging and Navigation: Extracting data from multiple pages by automatically moving through dynamic pages is very tedious. However, ParseHub provides a very easy and quick interface to automatically browse through dynamic pages to extract data.
- URL Crawling: ParseHub supports web crawling. For web crawling, Seed URL has to be provided to ParseHub.
- REST API and webhooks: ParseHub APIs support the programmatic management of the web scraping projects. Webhooks are available in ParseHub. The webhook provides the notification about the web scraping projects.
- CSV and JSON formats: ParseHub extracts unstructured data from the web and converts the same to a structured format with the option to specify the output format. CSV and JSON output formats are supported by ParseHub
- Google Sheets: In ParseHub, extracted data from the web can be imported to google sheets instead of downloading to excel or CSV format. The CSV or excel format can be downloaded each time the extraction is completed. But by using Google sheet, no need to download the date each time. Instead, ParseHub will update the Google sheet after completing the same project.
- Dropbox integration: ParseHub provided Dropbox integration to store the downloaded data.
- Navigate between different websites: Using ParseHub, data can be extracted by navigating through various websites.

The complexity of web scraping is depended on the structure of web pages. The application and features of web scraping methods rely on the layout of web pages. These days almost all big organizations or data scientists rely on web scraping techniques to extract information from the internet for decision making or research. Using the web scraped data, the organizations explore the investment opportunities, conduct market research, and decides on new products.

Applications of web scraping include Competitor Research, Industry Insights, Lead Generation, Research Data Gathering, Financial Data. In Competitor Research, web scraping is used to extract data regarding the competitors to obtain an understanding of their operation strategies like product information pricing approach and marketing pattern. Industry Insights refers to the performance of any specific industry. In this, web scraping will be applied to extract industry articles, stock details, and price information to obtain the industry's performance. Organizations will use web scraping to generate leads for their business. They will access web directories to extract business information to identify any potential opportunities related to their industry. Data scientists and researchers rely on web scraping technologies to extract data for the purpose of various researches. Data from multiple websites and online big data repositories can be extracted using web scraping techniques for the purpose of research. Financial data like stock news, statements of income, and stock market information can be extracted using web scraping methods to analyze the financial status of industries.

However, the common question regarding Web Scraping is regarding the legality of web scraping. Generally, web scraping is legal if it satisfies some rules. While carrying out web scraping, few rules have to be followed. Before doing web scraping, the first rule is that the data being extracted is public. The owner of the data has made the data public. The second rule is that don’t extract data from web pages that are controlled through user account credentials. Finally, make sure that the website has not blocked web scraping using the robots.txt file (Perez 2019).

Using NLP techniques, data from the internet can be extracted quickly and can be used for various studies and applications. This study relies on NLP techniques using the Python NLTK package to extract industry requirements from the internet to support national vocational qualification development.

2.2 Web Data Extraction

Automatically extracting data from the web is one of the main interests of researchers. Web-Crawling and Web-Scraping are the two main techniques in Natural Language Processing for automatically extracting the information from the web. Though the crawling and scraping techniques have individual persistence, they are applied interchangeably and created confusion at the implementation time. Web crawling is the organized automatic browsing of web pages to read and store the full content of a website intended for indexing and archiving. In contrast, web scraping is the automated method of extracting specific information from web pages to convert it into a structured format (Massimino 2016).

A web crawler is a software program that automatically surfs the web pages using sitemap protocol. The web crawler is also referred to as bot or agent, or spider. Generally, the web crawling will start with a Seed URL, a group of Uniform Resource Locator. The main algorithms used for web crawling are Breadth-First Search Algorithm, Depth First Crawling Algorithm, Page Rank Algorithm, and Online Page Importance Calculation Algorithm (Kumar, Jain & Agrawal 2016). Also, the other major approaches in web crawling are Large Site First, URL Ordering, Batch page ranking, Partial page ranking, HTTP Get request and dynamic web pages, customized sitemap, and using filters.

The breadth-first search approach is used when the depthless parts include the objective in deeper trees, and it works on level by level. Initially, this approach will search in the root URL and the adjacent URLs at the same level. The search will stop if the desired URL is found. Else it will start searching the following levels until the searching URL is located. If the searching URL is not found, then it will generate a failure report. On the other hand, the Depth First approach will start from the root URL and move forward through the most left child URLs. This will be repeated using backtracking and will cover all unvisited URL nodes until the search objective is achieved. If the number of child URLs is limited, Depth First is an efficient approach to managing problems related to the search.

In the Page Ranking approach, the significance of a page is decided by the total count of citations or backlinks to the providing page. In Online Page Importance Computation, the webpage's importance is identified using the cash value assigned to it. In this method, web pages with high value will be downloaded by the crawler. However, the Online Page Importance Computation is time-consuming as it may download the same webpage multiple times, and using the crawling the large site first is more feasible than the online page importance computation (Kumar, Jain & Agrawal 2016).

Dhanith, Surendiran & Raja (2020) proposed a new web crawler method using word-embedding-based Recursive Neural Network to overcome the issue of listing nonrelated URLs in the existing crawlers for web search engines. They used Adagrad-optimized Skip Gram Negative Sampling to calculate the cosine similarity to predict the relevance of the website based on the keywords. For the study, web crawlers using SVM, VSM, BFS, NB, and ANN were implemented in Python3 using urllib and BeautifulSoap packages. The results of the study proved that the proposed system using

Adagrad-optimized Skip Gram Negative Sampling outperformed the traditional web search algorithms.

A systematic process of mining and merging any required contents of a website is generally termed as Web Scraping. In web scraping, software application agents impersonate humans' roles to interact with the web servers. This agent is called a Web Robot. The web robot will surf the required web pages to extract the requested data and organize it in a structured manner (Glez-Peña et al. 2013). Web Scraping is now widely applied in many areas like job searching, advertisement, the health sector, journalism, and many other fields (Diouf et al. 2019).

A robust tool used for collecting unstructured data from online sources and transfer the same to CSV, Excel, JSON files in a structured format is known as Web Harvesting or Web Scraping (Miranda 2020). To make accurate and effective decisions, organizations need to rely on real and truthful data. The web is one of the primary sources for any organization to obtain accurate data, and they depend on web scraping to extract data from online sources. Web scraping can help any business obtain industry insights, generate leads, analyze competitors, and manage the brand reputation. Getting industry insights means collecting information on the performance of companies in a specific industry. This will allow the business to plan on its investment selections and expansion strategy. Using web scraping, organizations will be able to generate business leads. For this, organizations will web scrap online dictionaries to extract the name, location, industry, and contact info of any potential customers to expand their business. Another advantage for business by using web scraping is the competitor analysis. For competitor analysis, organizations can apply web scraping to extract the marketing plan, pricing strategy, reputation, and other vital information of other organizations. This will help the organizations to adjust or improve their strategy to compete with their rivals. Finally, the organizations can rely on web scraping to analyze their brand reputation. For this, organizations can apply web scraping to extract information from social media about the comments and public opinion on its brand. Also, organizations can apply web scraping methods to extract the public review from various review sites and apply any damage control to correct any issues to improve the public opinion if required (Miranda 2020).

Two main categories of web scraping are screen scraping and application programming interface (API) scrapping. The screen scraping will extract data from the website's source code using regular expression matching or HTML Parser. Simultaneously, the API scraping is used when the website returns XML or JSON files based on the structured HTTP requests (Dogucu & Çetinkaya-Rundel 2020). According to Glez-Peña et al. (2013), a web scrapping agent or framework will have three common tasks, namely, Site Access, HTML Parsing, and Content Extraction and Output Building. In the site access task, using the HTTP protocol, the web robot will initiate communication with the required website. Next, the web robot will extract the data from the web pages using regular expression matching. If required, regular expression matching may be used along with additional logic. Finally, a structured representation of the extracted data, suitable for further study and storage, will be formed by the web robot.

As per Singrodia, Mitra & Paul (2019), web scraping softwares use three major classifications of web scraping. They are Syntactic Web Scraping, Semantic Web Scraping, and Computer Vision web page analyzing. Syntactic classification uses HTML, CSS, and other web languages parsing to extract data from the targeted website. This classification uses methodologies like Content Style Sheet selectors, XPath selectors, URI patterns, and visual selectors. Semantic classification deals with semantic data. Structures like Web Ontology Language and Resource Description Framework are used in semantic web scraping. Computer vision procedure and Machine Learning techniques are used for computer vision web data analysis.

The legality of web data extraction is discussed by Macapinlac (2019). According to him, the scrapping of public data does not come under hacking and doesn't bound to any legal obligations. On the other hand, information that required user credentials for access and private information, profile are not allowed to web scraped by third parties. While extracting data from the web, one should have clarity on the data being extracted. The data should be publicly available to anyone who access the web. A website's data can be considered as public if the data access is not controlled with any user account. It is also important to note that the data from a website should be extracted if it blocks the web extraction using the website's robots.txt.

There are many cases internationally regarding data ownership violations through web scraping. One such famous case is hiQ Labs versus LinkedIn. LinkedIn is a leading professional network having millions of professional profiles, and hiQLabs is a research firm supporting data analytics in people and workforce analytics. Using web scraping, hiQLabs extracted data from the professional profiles publicly available on LinkedIn for their research purposes. At one point, LinkedIn blocked the tools used by the hiQLabs for web scraping and suited a file against hiQLabs for accessing public profiles from LinkedIn, arguing that Computer Fraud and Abuse Act (CFAA) has been violated by hiQLabs. However, the final verdict was in support of hiQLabs, and the court stated that accessing public data will not violate the Computer Fraud and Abuse Act (Perez 2019).

2.3 Application of Web data Extraction in various industry

In this technology and the internet-dominated world, people are heavily dependent on the internet to interact with each other. Internet is the leading media of communication these days. Business people rely on the internet for identifying industry requirements, business leads, and competitor monitoring (Drake, Thornock & Twedt 2017). Social media is an online platform where individuals interact with each other and create networks with people or groups with the same personal or professional interests (Wikipedia 2018). Social media platforms like Twitter, Facebook, LinkedIn.. etc., are data gold mines. Researchers and organizations extract data from social media for research or business purposes. Some of the main areas where the social media data were used are identifying public opinions, stock market prediction, market analysis, identifying and reaching customers, brand management, sentiment analysis, the performance of organizations and industries, etc. There are many types of researches are going on regarding the extraction of information from the internet, especially social media data. Many anonymous information can be identified by extracting and analyzing the data from the internet.

A Performance Measurement System was theoretically developed by Agostino & Sidorova (2016) to identify the impact of social media in business in terms of users' opinions and conversations, network structure, financials, and interactions. They conducted the conceptual study by implementing the qualitative methodology of literature review to create a primary view of social media measurement framed for the Performance Measurement System. To monitor organizational decision-making, learning and support, provide external accountability, and motivate individuals, public, private, and no-profit organizations have implemented the Performance Measurement System. They explored the metrics available for measuring the social media contributions to the business and the methods required to achieve the metrics.

According to their study, the financial indicators in the Performance Measurement System give an imitated evaluation of the social media contribution towards the financial aspects based on the social media investments. This financial factor of the Performance Measurement System, extracted from the social media platforms, provides the Return of Investment related to social media. Details of social media networks can be extracted using web scraping. This networking factor of the

Performance Measurement System will provide an insight into the social media users' network's contributions. The extracted details of the social media users' interaction can be used to identify the user activities. This will support the interaction factor of the Performance Measurement System. An organizations' ability to interact with social media users can be extracted using various web scraping methods, which will contribute to the Engagement factor of the Performance Measurement System. Thus the Performance Measurement System framework provides a clear picture of the impact of social media on the organization's performance (Agostino & Sidorova 2016).

Han & Anderson (2021) provided a simple method to extract online reviews and prices from multiple travel facilitating websites. They used Python libraries to implement a scraping mechanism. Their research provided a simple tool for hospitality researchers to quickly gather extremely valuable secondary data from the web resources. The application of web scrapping in the food industry was examined by Hillen (2019). According to their study, web scraping techniques can easily access real-time data related to food products, price, store, etc., for the purpose of agriculture and food economic research. The application of web data extraction in the new media was explained by Sundaramoorthy, Durga & Nagadarshini (2017). They introduced a new platform, called NewsOne, using web crawling and web scraping methods to extract the latest news updates from various national and international news portals and show them shortly and clearly to the customers. The web crawling techniques can extract text from retail websites and replace the manually collected commercial data to identify relevant customers for effective marketing(D'Haen et al. 2016).

Ulbricht (2020) points out the application of web scraping in politics by introducing a new concept called "Demos Scraping". Demo scraping is the approach of identifying public views by scraping information from social media to improve democratic decisions. Another practical implementation of web scraping is in the area of Stock Marketing. Maurya et al. (2019) used web scraping to extract the data from multiple stock market websites and used Random Forest Classifications and Regression algorithm to predict the future rates. A meta-crawling algorithm was used to access and extract Geospatial information from the web to enable the automatic detection of distributed geospatial resources (Li, Wang & Bhatia 2016).

Retail shops can increase revenue by improving their pricing strategy with the help of web scraping methods to extract the prices from their competitors. Jorge et al. (2020) demonstrated the application of web scraping in increasing the sales of a wine shop by analyzing the competitor's price. In another study, the Green Building Material Information was extracted using the web crawling technology and used ontology techniques to classify the extracted data to identify the certified Green Building Materials suitable for specific construction purposes (Hong, Lee & Yu 2019). Ayyappan & Matilda (2020) showed how to use web scraping in police investigations. They proposed a system of identifying the criminals and missing children by extracting photos from the web and comparing the same with a given image using face mapping.

To help recruiters to assess the candidates without a formal degree Capiluppi, Serebrenik & Singer (2013) proposed a mechanism to extract the individuals' activity in the social media and online technology platforms like LinkedIn, Facebook, GitHub, Stack Overflow ...etc., to analyze individual's professional/ technical reputation, reference signals, recommendations, and experience.

The web data extraction can be applied to different areas in the education sector. Natural Language Processing (NLP) application in the education sector started very recently (Abdous & He 2011). Using NLP's information extraction techniques, academics and scholars can easily avail more volume of online resources after sorting out unrequired information. Proper implementation of NLP will contribute to the institutional effectiveness by improving academic writing, learning objective formulation, assessment preparation, questionnaire generation, and subject information extraction (Alhawiti 2014).

Frazier, Davis & Vickery (2020) discuss the application of web scrapping using Python to produce NC State University associated editor data from various publisher's web sits. They used these harvested data from publishers' websites to generate a complete evaluation strategy for the University's journal investments. In another study, Anglin (2019) explains the Web Crawling and Web Scraping methods to collect data from various District policies from the different websites to support the education policy development.

The main disadvantage of using web data for the research, pointed out by the researchers in various industries, is the unavailability of historical data and transactional data.

2.4 Application of Extracted Job Demands

The accuracy of official labor market data analysis can be increased by creating and curating real-time knowledge from the online labor market information. As free-to-access employment details are available to the public, the time taken to access, analyze and conclude industry labor information has been reduced significantly. Researchers can rely on online job information to extract the industry's most demanded positions and the soft & hard skills required to perform in each position. The main stakeholders who are really interested in the online job data are organizations and companies, employees, educational institutions, and policymakers.

Manually accessing and analyzing the industrys' workforce demand is a tedious task. These days, employers from different industries and job seekers rely on online platforms for employee and employment search (Boselli et al. 2018). The dynamics of industry workforce requirements can be obtained by analyzing the employment opportunity data posted by the various industrys' organizations. The jobs posted on the online job portals represent the industry requirements (Johnson, Albizri & Jain 2020), (Dadzie et al. 2018), (Lovaglio et al. 2018). However, the concern is whether the jobs posted on the web represent the actual workforce demand of the industry.

Lovaglio, Mezzanzanica & Colombo (2020) conducted a study to compare the job data available on the web and the actual job requirements for the industries. The study included the time series comparison of the jobs posted on the Italian online job portals and the exact vacancies from the national statistics identified using the traditional labor market surveys. The jobs posted on newspaper websites, online job boards, employment agencies websites are extracted and stored for this study. The official workforce requirements in Italy were collected from the Quarterly survey conducted by National Statistics Office. For comparing the jobs extracted from the web and the actual skill demand data from the National Statistics Office, the job title in the bot dataset were classified based on the International Standard Classification of Occupation (ISCO) taxonomy using the supervised machine learning text classification. This time series comparison between the online job data and the employment data from the National Statistics Office showed that the workforce demands available on the web at a particular time period matched with the workforce demands available in the industries.

The organizations use the online labor data to recruit employees or understand the skills to identify the new trends in required positions. The employees or job seekers rely on the digital job information to identify the unique skillsets requirements for employment and upgrade their skills. The education institutions, especially the universities or education regulators, rely on digital employment data to determine the skill sets required for various industries and the job market trend to create or upgrade the curriculum. Also, academic researchers rely on online job data for different research purposes. By analyzing the web job market data, the policymakers can decide on the training processes and make any improvements if required. The main goals of online job market analysis are skill analysis, identifying candidates for jobs, forming teams, detecting (Papoutsoglou et al. 2019).

According to Boselli et al. (2018), the web is the primary resource for both job seekers and employers to satisfy their requirements. Online job portals are the primary source of employment demand and supply matching. The new job openings are published in the job portals by the employers, and the job seekers access the job portals to apply to the posted job openings. To obtain up-to-date information about job market dynamics, it is necessary to analyze the job requirements frequently. Information that can be extracted from the online job data are title and full description. The title provides a short summary of the position, and the full description includes detailed information about the position, like the required skill sets and competency to perform the tasks related to the post.   One of the reliable sources in getting the job market is the online job portals, and it supports evidence-based decision-making for industries.

Lo et al. (2020) argued that the job opening indicates an organizations' operation performance in terms of financial ratio and stock returns. To prove this, they web scraped job postings of various publicly listed Taiwanese organizations from the job portal www.104.com.tw for three months. Through web scraping, they extracted 578,965 vacancies from 830 companies. The regression results' coefficient was significantly positive, showing that an organization's performance is positively connected with the rate of published jobs. Pazmiño & Diaz (2018) conducted a research to identify the trend of workforce demand to satisfy the requirements of industries and determine the provinces of Ecuador where the labor demand is high. The jobs posted along with the required skill sets from the online job portals were extracted for the study. The extracted data provided precise analytics of the required workforce per province and industry sector. This study becomes a basis for making economic policies and workforce distribution strategies in Ecuador.

The comparison of skill requirements of the industry with the available skills produced by the providers like technical or higher education will help both job seekers, employers, and government institutions (Szabó 2011). By this comparison job, seekers can upgrade their skills to secure employment, and employers can plan for employee recruitment, and government institutions can plan for workforce preparation planning and national education policies. Szabó (2011) proposed an ontology-based comparison of job roles and learning outcomes to plan for the workforce and training. The author presented a methodology of extracting job lists with required skills from online job portals, preparing job role ontology by identifying the job role requirements from the extracted job skill sets, and comparing the job role ontology with the learning outcomes ontology to identifying the gap between the employer demand and employee supply.

Litecky et al. (2008) carried out a study in the US to map the required skillset with the IT job description. They extracted quarter-million IT employment details from various job portals like HotJobs.com, simplyhired.com, and monster.com. Using the NLP techniques, relevant skill set information for each position was identified from the extracted data. Different research was carried out in Europe using WoLMIS to classify multilingual employment advertisements published in multiple job portals, based on the International Standard Classification of Occupations (ISCO) taxonomy (Boselli et al. 2018). The intention of the study was to assist the European Vocational Education and Training Authority in identifying the job market dynamics and trends in various countries regardless of the language difference. Around 1.8 million employment openings posted in different European countries were extracted to identify high-demand skills and assist the European

Vocational Education and Training Authority in understanding the skills required for various industries.

Lovaglio et al. (2018) analyzed 276,909 ICT and Statistical employment opportunities advertisement extracted from Italian job portals to identify the skills required to satisfy the position requirements. They implemented many machine learning techniques to determine the best matching skills compared to the ISCO occupations. N-gram technique was used to identify specific skills and were confirmed by the domain experts. Classification analysis using the Support vector machines was used for the skill set classifications. The classification results showed an accuracy of 0.765.

Mauro et al. (2018) proposed a semi-automated methodology using machine learning algorithms and expert judgment to help organizations implement effective strategies that support the recruitment and development of a qualified workforce. Their study considered the Big Data job families to identify the required skill sets and competencies. Big-Data-related employment opportunities with job titles, descriptions with required skillsets and competence, geographic information were extracted from various online job portals using web scraping. Using the basic natural language processing techniques and experts' judgment, different job families related to Big Data were identified. The required skill sets for each of the Big Data job families were identified using the Mixed Membership model Latent Dirichlet Allocation. They extracted the job details from various US-based online job portals. They recognized that Business Analysts, Data Scientists, Big Data Developers, and Big Data Engineers are the four main categories related to Big Data. The study also identified the required skill sets and competencies needed for each type for helping the organizations’ human resource team to plan the recruitment and skill development related to Big Data.

Dadzie et al. (2018) used job summary information, crawled from LinkedIn to analyze the skill sets required for the Data Scientists jobs across Europe. Their study accessed the demand from the job market, and analyzed the same to help the training providers and job seekers to fill the gap between their existing capabilities and skill requirements. The study examined the problems like the demand for a data scientist in European Union, the level to which the demand matched the current workforce skillsets, and the support provided to the existing workforce to empower their skills according to the demand. The study did a detailed analysis of the data related to current job demands from various sources to answer these problems. The main drawback of this research was that it didn't discuss the result validation.

Ontology is the shared knowledge of any particular domain formed as a collection of examples, tasks, ideas, associations, axioms. To provide clear information of any targeted field, Ontology explains the terminologies of a domain. An ontology-based knowledge base was developed to identify the job demands and provide real-time information on industry requirements (Thi et al. 2020). In-depth knowledge, ideas, understanding, and connections of the targeted domain are required to build Ontologies. The development of ontology involves capturing information from various sources and mapping them to the ontology instances. The ontology developed for processing the job demands included knowledge extracted from the online job portals. This ontology represented the capability, knowledge, and skills required to demonstrate the competency to perform any given task. 350 job posting from various national and international online job portals were extracted for the ontology creation. A rule-based approach was implemented to capture information and populate the relevant ontology concept instances. This job skill demand ontology helped identify the required industry skill demand quickly. F1 score was used to measure the effectiveness of the extraction, and it showed an accuracy with an average of .83.

Johnson, Albizri & Jain (2020) implemented a framework to develop a curriculum for a master’s program in business analytics by identifying the industry’s requirements in terms of concepts, skills, knowledge, and tools. They scraped job descriptions related to Business Analyst positions from the online job portal www.indeed.com to extract the required skills, knowledge, and tools related to the Business Analyst position in the Northeast Region of the United States. Using the BeautifulSoup and Selenium packages of Python, Pillai & Amin (2020) web scraped 10 Indian IT companies’ employment details. Their objective was to identify the top skills required by the IT companies, help the graduates prepare for the job market, and support the universities to update their curriculum. The technical aspects of the analysis were not covered in the research.

The majority of these studies used Accuracy and/or F-measure to identify the proposed system’s reliability or algorithms.

3. Research Methodology

All data used in this study are publicly available online, so data confidentiality is not a concern for this study. The data collection methods used for this research don't include any participants other than the researcher. The data or data collection methods don't violate any research ethics. The data for this study was extracted from the online job portal https://ae.indeed.com/. The job information published by the employers ate publicly available on the portal https://ae.indeed.com. According to Macapinlac (2019) and (Perez 2019), the scrapping of public data does not come under hacking and doesn't bound to any legal obligations.

The number of employers using the web to advertise the job opening is increasing significantly. Employers publish the job titles and required skill sets in the web job portals. These data, publicly available on the web, represent the current labor market dynamics and support various stakeholders like private & government institutions, researchers, and analysts in decision-making (Lovaglio et al. 2018). This research aims to establish Natural Language Processing techniques to ease the qualification development process by identifying the industry's employment requirements.

The theoretical framework of this research consists of 6 steps.

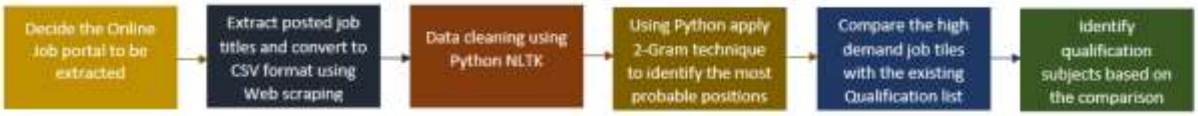

Decide the Online Job portal to be extracted
Extract posted job titles and convert to CSV format using Web scraping
Data cleaning using Python NLTK
Using Python apply 2-Gram technique to identify the most probable positions
Compare the high demand job tiles with the existing Qualification list
Identify qualification subjects based on the comparison

Figure 1: Theoretical Framework

1. Identify a suitable job portal to extract industry requirements
2. Extract the unstructured employment data from the identified web portal and convert it to a structured format
3. Apply Python NLTK techniques to clean the extracted data

4. Identify the most probable positions using the 2-Gram technique
5. Compare the high demand job titles with the existing National Qualifications
6. Recommend new qualifications

Identify a suitable job portal to extract industry requirements

The first step in the process was to identify a suitable online job portal for the study. www.Indeed.com is one of the top job portals in the World. This site uses the web crawling technique to extract posted job requirements from different companies and various job portals worldwide. Indeed.com's UAE job portal, https://ae.indeed.com/, was used in this research for extracting the industry's employment requirements in UAE.

The main reasons for using the https://ae.indeed.com/ job portal for this study are:

1. Indeed.com is one of the leading job portals in the UAE
2. It includes web crawled data (job opportunity data) from other major job portals and employers in UAE. This will help the research accessing a wide range of employment advertisements published on the web from all across the UAE.
3. The information structure used in this website is advantageous in accessing the data quickly using web scraping.

Extract the unstructured employment data from the identified web portal and convert it to a structured format

The next step is to extract the data from the identified job portal. Web data can be extracted using the BeautifulSoup package of Python. However, the web scraping coding may take more time for dynamic websites like https://ae.indeed.com. It will be time-consuming to understand the HTML, CSS structure of the site before extraction. Also, the primary bottleneck will be accessing data from multiple pages of dynamic websites. Moreover, IP rotation may be required if the target website has a Web Scraper blocker. Considering these factors, it was decided to use a free web scraper tool to extract the employment data from the web.

This study used the ParseHub, a free and powerful web scraper tool, to extract employment data from the job portal https://ae.indeed.com/. The advantage of using ParseHub is that it is effortless to extract data from any website, especially for data spread over multiple pages, and have complex structures implemented using javascript or CSS. The main features of ParseHub are Scheduled Run, Automatic IP Rotation, Ajax and Javascript Enabled interactive websites, Expressions and Conditionals, CSS, XPATH, RegEx Selectors, Tables, and maps, Extract HTML/text attributes, Download images/files, Access Secured web data, Scrolling pages, Forms and inputs, Dropdowns/tabs/pop-ups, Paging and Navigation, URL Crawling, REST API and webhooks, CSV and JSON formats, Google Sheets, Dropbox integration, and Navigate between different websites.

UAE's IT industry employment data were considered for this study. The data from https://ae.indeed.com/IT-jobs-in-UAE was extracted using the ParseHub (Figure 2). The job listing page has mainly two pieces of information: job title and job description. The job title contained a short description of the position with not more than ten words, and the job description included details of the position, including the required skill sets and competency to perform the job tasks. The main objective of this study is to identify the qualifications needed for the selected industry, so it was decided to access only the job titles. After loading the page in the ParseHub, the job titles were highlighted as the field to extract. The IT job list had more than 90 pages.

To have a recently posted dataset not older than two months, it was decided to extract the data from the first 34 pages only (Figure 3). The ParseHub extracted a total of 284 IT job titles from the job portal, and the extracted data was downloaded in CSV format. In the downloaded CSV file, data from each page was listed under separate rows, as shown in Figure 4. After downloading the data, the first task was to arrange the job titles in one column. Excel's TRIM(CONCATENATE()) functions combination was used to place the data in a single column named 'Job_Title'.

Apply Python NLTK techniques to clean the extracted data

The third step in this research framework is to apply Python NLTK techniques to clean the extracted data. The job titles included a short description of the position with less than ten words, and the data required cleaning as included many special characters and junk words. The next step was to clean the data by removing special characters and stop words used in the job titles. Since the data is in an unstructured format, it needs to be preprocessed to clean and standardize the text and remove any noise. Noise is a portion of text that is not appropriate to the context and the expected output.

The noise removal includes the exclusion of stop words like 'the', 'is', 'am', 'of', 'on', 'to' ..etc., URLs, hashtags, industry-specific words, and punctuation. After removing the noise, the text can be normalized using Stemming or Normalization. The process of removing the suffixes like "ing", 'ly", "es"…etc from a word using a rule-based process is known as Stemming. In contrast, the structured step-by-step method of extracting the root form of a word is Lemmatization.

The job titles were cleaned using the Python NLTK package. A function title_clean() was created for this purpose, and the cleaned job titles were stored in a new column, 'Cleaned_Job_Title' (Figure 5). While posting the jobs, the Job titles included many meaningless words, which negatively impacted job title analysis. To overcome this, NLTK's WordNetLemmatizer function was used to lemmatize the cleaned job titles (Figure 6).

Identify the most probable positions using the 2-Gram technique

After cleaning the job titles, the next step is to identify the most repetitive job titles. This step is to identify the industry demands. After lemmatization of the job title, NLTK's Bi-Gram technique was applied to identify the top 10 job titles (Figure 7). The top ten job titles in the IT industry included project manager, customer service, data analyst, service manage, system administrator, customer service, and information technology. This implies that the IT industry requires a workforce that can satisfy the requirements of these positions. To supply these workforce demands of the IT sector, the UAE vocational education sector shall offer qualifications related to these job titles.

Compare the high demand job titles with the existing National Qualifications

In the previous step, by extracting the job postings information from the online job portal, the UAE IT industry's top 10 required job titles have been identified. The next step is to determine whether any qualification related to the identified job titles is currently existing. To perform this task, a list of existing qualifications has to be obtained.

According to the NQA (2012), the UAE has 12 industry sectors. The twelve industry sectors in UAE are:

1. Government services and public administration
2. Community, health and social services
3. Business, administration and financial services
4. Tourism, hospitality, retail and leisure services including personal care services
5. Arts, culture and entertainment
6. Education, learning and social development
7. Building and construction, estates and assets development and management
8. Utilities and infrastructure
9. Energy resources - oil, natural gas, petrochemical, chemical and mining/quarrying
10. Manufacturing
11. Logistics and transport
12. Agriculture, livestock and fishery

Each of these 12 sectors has subsectors, and the IT field is categorized under the sub-sector Utilities Supports in the sector ‘Utilities and infrastructure’. Qualifications developed by the NQA are categorized under these 12 industry sectors. The list of NQA qualifications by sector is published on the website of the Awarding Body ACTVET. Since the list is available on a single page, the qualification list was extracted manually from the website and converted to excel format. Using simple search and VLOOKUP functionality of excel, the existing qualifications and the identified top 10 job titles were compared to recognize qualifications related to the job titles.

Recommend new qualifications

Now the final step in the research framework is to recommend qualifications for the selected industry. In the previous step, by comparing the existing qualifications and the identified top 10 job titles, the currently endorsed qualifications related to the specified job titles are recognized. This comparison revealed that out of the top 10 IT job titles, the UAE vocational education sector has qualification-related to only one job title: Customer Service. This implies that the UAE IT sector needs qualifications related to project management, data analyst, service manager, system administration, and information technology. And based on this observation, five qualifications can be recommended for the UAE IT industry.

This study uses a semi-automated approach to recommend qualifications for the UAE industry by implementing web extraction techniques using Natural Language Processing.

4. Results and Discussion

As stated before, this study's ultimate goal is to investigate the application of Natural Language Processing techniques to simplify the qualification development process by automatically detecting the industry's employment requirements. Using the web scraping techniques, employment data posted on the job portal was extracted, and the top 10 job titles in the UAE's IT industry were identified (Figure 8).

The next step is to identify if any NQA endorsed qualifications related to the identified job titles' subject area are available. As per the Qualifications Framework for the Emirates Handbook (2012), there are 12 industry sectors in UAE. The National Vocational Qualifications are developed based on the requirements from these 12 industry sectors. The NQA endorsed vocational qualifications are categorized by the industry sectors. The Abu Dhabi Centre for Technical and Vocational Education and Training (ACTVET), an awarding body under the NQA, publishes the list of national qualifications, by sector, on their website. The list can be accessed from the URL https://www.actvet.gov.ae/en/AboutUs/Structure/QualificationsDepartment/Pages/default.aspx. Each qualification The ACTVET regularly updates this list based on the qualifications endorsed by the NQA.

According to the current qualification list, there are a total of 49 national qualifications endorsed by the NQA. 18 qualification under 'Utilities and infrastructure', 11 qualifications under the sector 'Business, administration and financial services', five qualifications under the sector 'Tourism, hospitality, retail and leisure services including personal care services', four qualifications under the sector 'Community, health and social services', three qualifications under the sector 'Education, learning and social development', another three qualifications under the sector 'Arts, culture and entertainment’, two qualifications each under the sectors ‘Government services and public administration’ and ‘Building and construction, estates and assets development and management’, and one qualification under the sector ‘Energy resources - oil, natural gas, petrochemical, chemical and mining/quarrying’.

As per the UAE’s sector and sub-sector categorization, the Information Technology is under the sub-sector Utilities Supports in the sector ‘Utilities and infrastructure’. According to the qualification list, there are 18 qualifications under the sector ‘Utilities and infrastructure’. However, all the 18 qualifications under the Utilities and infrastructure are related to Mechanical or Electrical engineering, and not a single qualification was associated with Information Technology.

Though there were no national qualifications related to Information Technology, it was observed that the sector ‘Tourism, hospitality, retail and leisure services including personal care services’ has a qualification named ‘Certificate 4 in Customer Service’. And, the qualification ‘Certificate 4 in Customer Service’ has some units related to Information Technology. So, it was decided to consider ‘Certificate 4 in Customer Service’ as a matching qualification for the IT customer service.

To identify the required qualification, the listed qualifications were extracted manually from the ACTVET website and compared with the automatically extracted job titles from the online job portal. This high-level comparison was made manually at the level of qualification title and position title. Necessary Skillsets attached to the positions were not considered for identifying the related qualification. By comparing the identified job titles with the existing list of endorsed qualifications, it is recognized that the qualification ‘Certificate 4 in Customer Service’ is matching with the titles ‘Customer Service’ and ‘Customer Specialist’, and the remaining positions don’t have any related qualifications. This shows that the Information Technology industry has a high demand for the positions like Project Manager, Customer Service, Data Analyst, Service Manager, SystemAdministrator, Service Analyst, and Customer Specialist.

But only two titles, 'Customer Specialist' and 'Customer Service' have related qualifications. The remaining titles Project Manager, Data Analyst, Service Manager, System Administrator, and Service Analyst don't have any qualifications that support the related workforce generation. This clearly implies that the Information Technology sector in the UAE needs vocational qualifications to support the high workforce demand for Project Manager, Data Analyst, Service Manager / Analyst, and IT specialist (Table 1).

Industry	Job Title	Related Qualification
IT	Project Manager	Not Available
IT	Customer Service	Certificate 4 in Customer Service
IT	Data Analyst	Not Available
IT	Service Manager	Not Available
IT	System Administrator	Not Available
IT	Service Analyst	Not Available
IT	Customer Specialist	Certificate 4 in Customer Service
IT	Information Technology	Not Available

Table 1: Job Title and Qualification comparison

Discussion

Vocational education can significantly reduce the gaps between the industry demand and workforce supply and reduce the unemployment rate. One of the critical factors in maintaining a country's 'Internationally Competent' status is the availability of relevant qualifications that empower the citizens with skills and knowledge to cope-up with the highly evolving system, resources, workplace environments, and technologies. It will significantly improve the workforce quality and productivity. Also, this will enhance the country's well-being by producing excellent cost management and workforce management systems. By providing suitable vocational education qualifications, issues related to the scarcity of skilled laborers pertaining to the specialized and technical sectors can be managed and can significantly reduce the unemployment rate. By producing a workforce with the required skills and knowledge training, the technical and specialized positions in the industries can be filled quickly. The availability of qualifications that satisfy the specialized requirements of the industry will provide the existing and potential staff with the opportunity to enhance their skills and help them stay in employment for a long time.

Employers prefer vocational education graduates as they are well prepared with ready-to-work skills, and no further training is required. However, the employers' expectations are not satisfied. There is a big gap between the industry requirements and the skills obtained by the graduates in UAE. To manage the scarcity of skilled employees, the UAE government promotes vocational education among the nationals and gives high priority to establishing high-quality standards for the vocational education system. Setting high standards for the vocational education system is a cyclic process, and it starts from the identification, planning, and development of qualifications. Usability, consistency in the outcomes, return on investments, stakeholder satisfaction, and continuous improvement are the few factors that decide the quality of a qualification.

Government policies and strategy, financial sustainability, provisions, and industry requirements are the factors that decide the necessity for developing a qualification. Capturing industry requirements is the main element in the qualification development and quality assurance processes. The success of a qualification relies on how well the industry requirements are captured and the extent to which the qualification satisfies the industry demands. The qualification development processes implemented by the National Qualification Authority (NQA) in UAE include a high standard and integrity process. The current requirements and demands shown in each industry sector are captured to develop any vocational qualifications. To capture the industry demands and conditions related to the workforce, the NQA constantly liaises with the industry representatives and works closely with the industry team to develop multiple vocational qualifications. Web Scrapping methods can be successfully implemented to radically reduce the time and cost involved in the process of identifying and updating the industry requirements.

Web data is the primary resource for the researchers for data collection. Many studies are conducted in the area of Web data extraction. Web data extraction can be applied to any industry. Almost all industries have relied on and still depending on web data for their process enhancements. Various industries have implemented web data extraction techniques for their operational purposes and research. One of the main areas where the web extraction was applied was the extraction of employment details from the company websites and online job portals. From the literature review, it was evident that the job details published on the web will help in analyzing the skills demands for each industry. The expensive and time-consuming element in vocational qualification development is the identification of industry requirements. The proposed semi-automated technique in this paper is an effective replacement for the manual requirement gathering element in the qualification development process. The industry requirements were identified by extracting the jobs posted on the online job portals.

By applying the NLP techniques like web scraping, data cleaning, lemmatization, and n-Gram, the IT industry’s advertised job titles were easily extracted from the online job portals, and analyzed to identify the top 10 IT job positions. The IT industry in the UAE is included in the sector ‘Utilities and infrastructure.’ By comparing the extracted top 10 Job titles with the qualifications under the category ‘Utilities and infrastructure’, published on the ACTVET website, this study was able to identify and recommend the required qualifications for the IT industry.

5. Conclusion

Vocational Education helps the industries in preparing up to date and ready to start workforce. The effectiveness of vocational education can be measured based on its contribution to the industries. Measuring the vocational education impact on industries is a cyclic process, and it should start at the stages of qualification identification, planning, and development. Identifying the need for a qualification or identifying required qualifications for an industry is the first step in developing a qualification. For this, authorities responsible for developing qualifications must be monitoring industry developments and requirements to identify the skills demands. And based on the identified needs, the authorities shall propose & initiate any required qualifications proactively to avoid any shortage of workforces.

Recognizing and updating the requirements of industries at regular intervals is a costly and time-consuming process. Usually, the qualification development team uses surveys and face-to-face meetings with industry representatives to collect the industry demands. Various studies surveyed in this research showed that job requirements posted by employers are an essential and reliable source in identifying the industry requirements. Multiple studies proved that Natural Language Processing could be used successfully to extract and analyze data from the web.

Natural Language Processing uses computer systems to manage the analysis, processing, and understanding of human languages. Different aspects of human languages like sentence structure and meaning can be handled by using NLP. This covers the extraction of data from the web and other sources, text categorization, checking the spelling and grammar, machine translation, and prediction of words (Single, Schmidt & Denecke 2020). A popular high-level programming language called Python is widely used for Natural Language Processing. Python has many libraries that support the easy process of natural languages. One of the main Python libraries used to process and extract information from natural language is NLTK.

NTLK provides a very user-friendly interface to manage various corpora with libraries to support the text processing operations like parsing, classification, tagging, tokenization, and stemming. NTLK also supports the reasoning of semantics and context-based discussion of conversations. Operations like Part-of-speech tagging, Entity Extraction, Tokenization, Stemming, Parsing, Semantic Reasoning, and text classifications can be carried out using the NLTK libraries. NLTK processes languages by accepting strings as input and provide the output as a string or an array of strings.

This research conducted a comprehensive literature review and showed that web data extraction is widely applied in almost every industry. Web Scraping is a programmed strategy to extract data from the web. Web scrapers are used to extract the required information from a single website or a group of websites. Most of the data available on the web are in an unstructured format. Using various web scraping methods like Application Program Interface (API), web administrations, manual software coding, and ready-made applications, those unstructured data on the web can be extracted, organized, and then converted into a structured format and stored as CSV or JSON files.

Python provides multiple libraries like Requests, Beautiful Soup 4, lxml, Selenium, and Scrapy to implement web scraping quickly. However, accessing data from various pages of dynamic websites using manual scripting is not easy. Also, if any Web Scraping blocker is applied on the target website, then applying the IP rotation to overcome the blocker is not straightforward. So this study used a free and powerful web scraper tool called ParseHub to extract the employment data from the web.

The main contribution of this study is to improve the National Vocational Qualification development process by implementing a semi-automated technique to reduce the time taken in collecting the industry requirements. Many researchers used web scraping techniques to extract employment data from job portals for various studies.

This study was concentrated on the IT job sector to identify the high-demand jobs and then identify and recommend qualifications to satisfy the workforce requirements accordingly. The Information Technology employment data published on the UAE job portal, https://ae.indeed.com/IT-jobs-in-UAE, were extracted to determine the Information Technology industry requirements as part of improving the NQA qualification development process. Job title and job description were the two pieces of information listed on the online job portal. The job title contained a short description of the position with not more than ten words, and the job description included details of the position, including the required skill sets and competency to perform the job tasks. For this research, only the Job titles were considered and extracted the same using the web scraper tool ParseHub and the extracted data saved in CSV format. Various Natural Language Processing techniques were applied to the extracted employment data to obtain the top 10 job titles.

The list of 49 national qualifications endorsed by the National Qualifications Authority has been manually extracted from the website of the awarding body, Abu Dhabi Centre for Technical and Vocational Education and Training (ACTVET). The manually extracted qualifications and the automatically extracted job titles were compared to identify the existing and required qualifications. This high-level comparison was based on the job titles. The required skillsets attached to the positions were not considered for determining and recommending the qualifications needed for the Information Technology sector.

The extracted top job titles were compared with the existing endorsed vocational qualifications and identified the missing qualifications for the Information Technology industry. This study showed that the Information Technology sector in the UAE needs vocational qualifications to support the high workforce demand for Project Manager, Data Analyst, Service Manager / Analyst, and IT specialist. Overall, the proposed semi-automated system was capable of recognizing the industry's workforce demand and propose vocational qualifications to support the workforce development accordingly. This research also established that by implementing the web extraction technologies and various NLP methods, the industry requirement can be easily identified and can significantly improve the qualification development process.

6. Limitations and Future Work

A skillful and competent workforce is a vital part of improving any country's economic and social aspects. Industries should be able to employ expert employees and upgrade their existing staff's skills to cope with the fast-changing industry environments and technologies. From layman to higher management need to have current knowledge about the industry. So the main challenge in any industry is to obtain suitable and expert employees. However, scarcity of skilled laborers pertaining to the specialized and technical sectors is the major challenge that all countries face to keep the 'Internationally Competent' status. Vocational Education and Training is a solution for equipping the workforce with the required skills. The employers prefer vocational education graduates as they are well prepared with ready-to-work skills, and no further training is needed. Vocational education and training can considerably manage the gaps between the industry demand and workforce supply and reduce the unemployment rate.

Vocational education prepares people for the technical sector and makes them work as technicians in various industries. The vocational qualifications are developed based on the industry requirements. Ensuring the effectiveness of vocational education is a cyclic process, and it starts from the identification, planning, and development of qualifications. Quality of qualification is referred to its usability, consistency in the outcomes, return on investments, stakeholder satisfaction, and continuous improvement. A qualification's quality is identified by its contribution to the learners' skill development, increased employability, and progress in innovations and productivity. Capturing industry requirements is the main element in the qualification development and quality assurance processes. The success of a qualification relies on how well the industry requirements are captured and the extent to which the qualification satisfies the industry demands. The first step in developing a qualification is the process of identifying the need for the qualification. To provide the required staffing on time for each industry, it is mandatory to continuously monitor the industry requirements and skill demands. Regularly identifying and updating the industry requirements is a manual process and is expensive and time-consuming.

As a solution, this study proposed a semi-automated system to replace the manual requirement capturing process in the qualification development cycle. The industry requirements were identified by extracting the jobs posted on the online job portals. The proposed system used the NLP techniques like web scraping, data cleaning, lemmatization, and n-Gram to identify the high-demand job roles in the industry and compare the same with the existing qualifications to identify and recommend vocational qualifications to equip the workforce with the required skill sets.

However, regardless of the contributions made, this research has many limitations worth mentioning. This study proposed the semi-automated mechanism for the national qualification development based on the IT sector only. In the future, more industry sectors can be included to recommend suitable qualifications for the industry workforce enhancement.

The proposed semi-automated system recommended the required national qualifications for the industry based on the job tiles extracted from the online job portals. The proposed approach was based on the high-level analysis constructed on the job titles and recommended the qualification. Due to the time limitations, the skill sets and competencies needed to perform the job roles mentioned in the job description were not considered for this paper. Extracting and analyzing the skill sets and competencies could have provided an in-depth vision of the industry requirements. In the future, the job description will be extracted to analyze the industry requirements in detail.

This study didn’t cover the outcomes or performance criteria of the proposed qualifications. The next version of the study will identify the skill sets and competencies from the extracted job descriptions and recommend qualifications along with the required qualification outcomes and performance criteria.

In this recommended approach, the extraction of existing national qualifications from the awarding body website was done manually. Since very few qualifications were listed on the awarding body website, it was easy to extract the qualifications manually. Also, the comparison of identified top 10 job titles with the extracted qualifications was made manually. In the future, machine learning techniques can be applied to compare job titles and currently endorsed qualifications for recommending new qualifications and the required qualification outcomes and performance criteria.

Since the extracted data from https://ae.indeed.com/ have employment data from various job portals in the UAE, there are high chances of extracting duplicated job opportunities from more than one job portals. It was manually checked and confirmed that the data used in this study did not include any duplicated job titles from the same company. In the future, to avoid duplicate jobs, company information will also be extracted along with the job listings. This will help to sort out the duplicated job postings from the same company automatically.

7. Appendix

Figure 2: Parsehub – Web data extraction

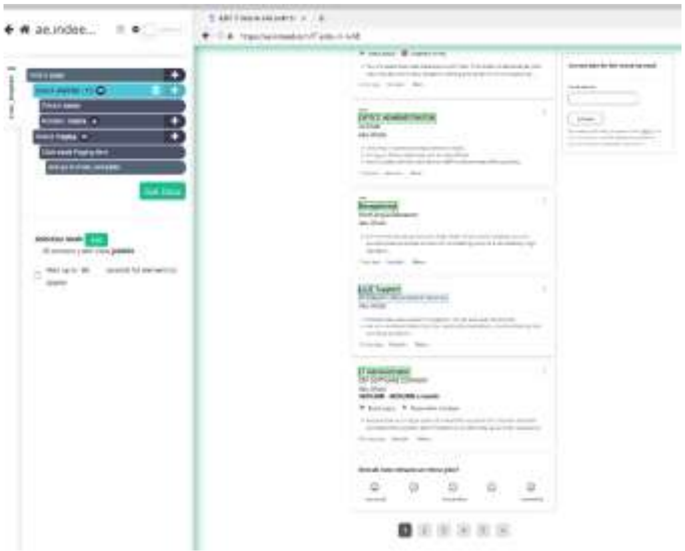
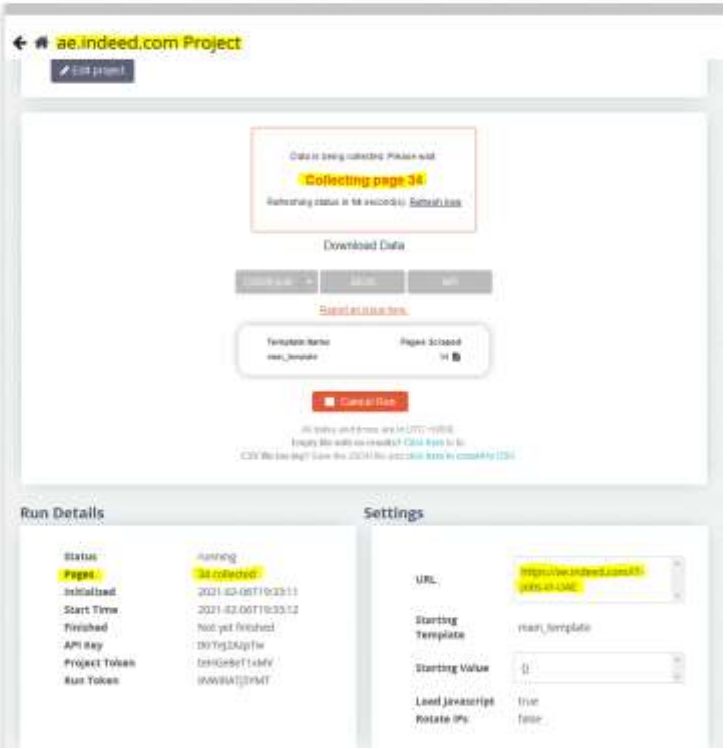


Figure 3: Parsehub – Wed data extraction result summary


ae.indeed.com Project

Edit project

Data is being collected. Please wait.

Collecting page 34

Refreshing status in [illegible] second(s). Refresh now

Download Data

[illegible]

Template Name	Pages Scraped
main_template	34

Cancel Run

Run Details

Status	running
Pages	34 collected
Initialized	2021-02-06T10:33:11
Start Time	2021-02-06T10:33:12
Finished	Not yet finished
API Key	[illegible]
Project Token	[illegible]
Run Token	[illegible]

Settings

URL	https://ae.indeed.com/[illegible]-jobs-in-UAE
Starting Template	main_template
Starting Value	{}
Load Javascript	true
Rotate IPs	false

Figure 4: Parsehub – Extracted Data Structure

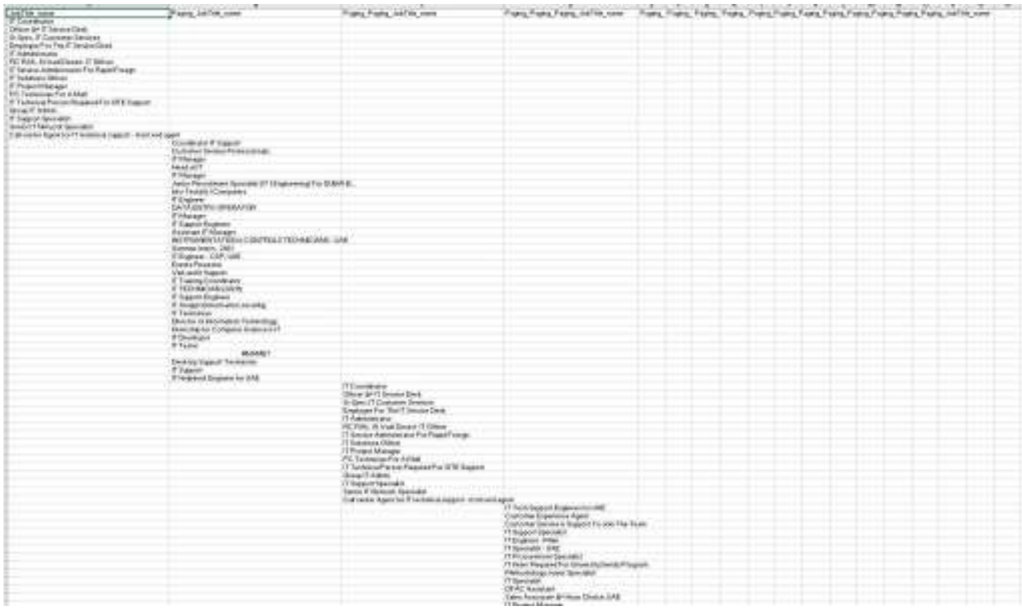
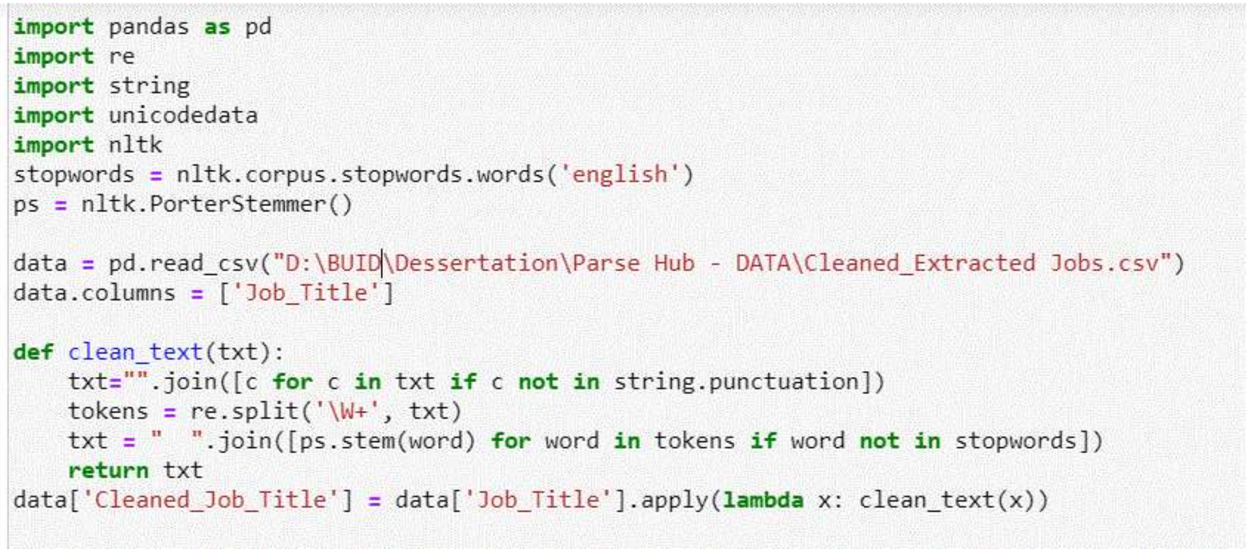
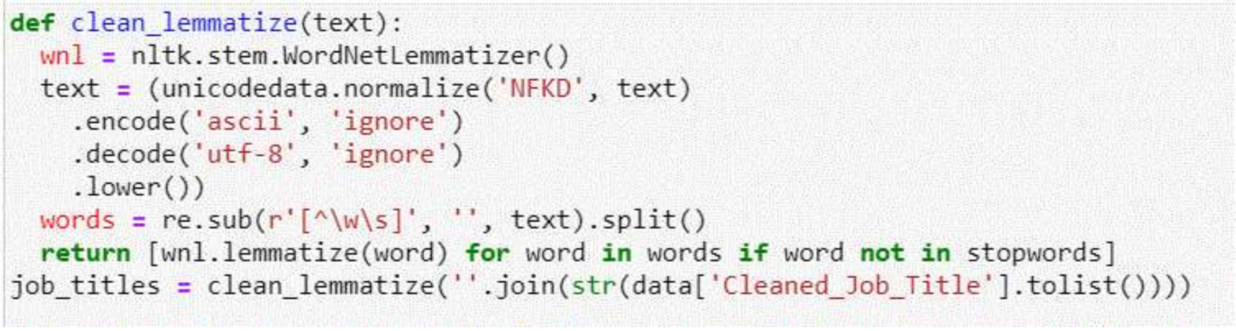


Figure 5: Python code for job title cleaning



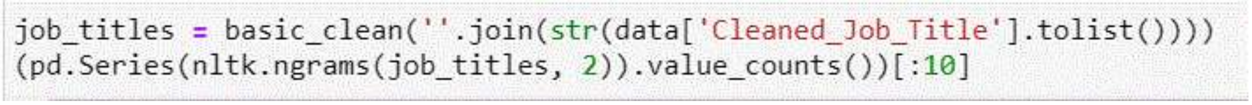
```
import pandas as pd
import re
import string
import unicodedata
import nltk
stopwords = nltk.corpus.stopwords.words('english')
ps = nltk.PorterStemmer()

data = pd.read_csv("D:\BUID\Dessertation\Parse Hub - DATA\Cleaned_Extracted Jobs.csv")
data.columns = ['Job_Title']

def clean_text(txt):
    txt="".join([c for c in txt if c not in string.punctuation])
    tokens = re.split('\W+', txt)
    txt = "  ".join([ps.stem(word) for word in tokens if word not in stopwords])
    return txt
data['Cleaned_Job_Title'] = data['Job_Title'].apply(lambda x: clean_text(x))
```

Figure 6: Python Lemmatizer code



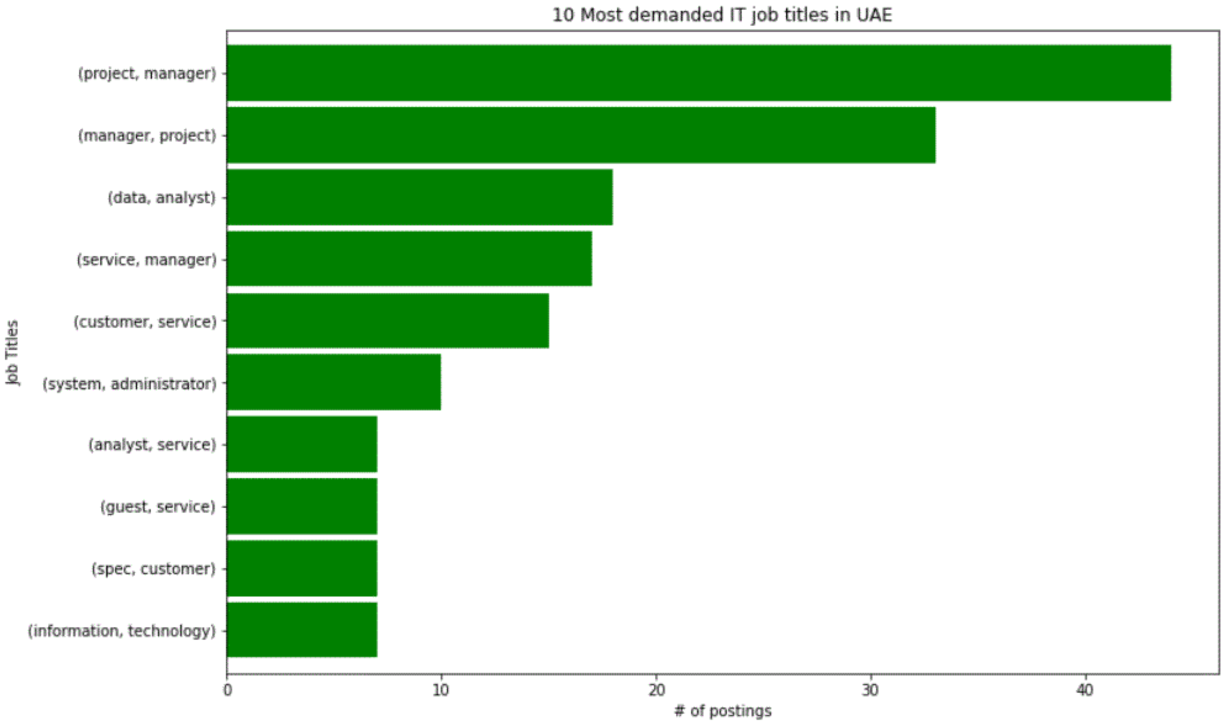
```
def clean_lemmatize(text):
  wnl = nltk.stem.WordNetLemmatizer()
  text = (unicodedata.normalize('NFKD', text)
    .encode('ascii', 'ignore')
    .decode('utf-8', 'ignore')
    .lower())
  words = re.sub(r'[^\w\s]', '', text).split()
  return [wnl.lemmatize(word) for word in words if word not in stopwords]
job_titles = clean_lemmatize(''.join(str(data['Cleaned_Job_Title'].tolist())))
```

Figure 7: Python Bi-Gram code



```
job_titles = basic_clean(''.join(str(data['Cleaned_Job_Title'].tolist())))
(pd.Series(nltk.ngrams(job_titles, 2)).value_counts())[:10]
```

Figure 8: Top 10 IT jobs in UAE



8. References

Abdous, M. & He, W. (2011). Using text mining to uncover students' technology-related problems in live video streaming. *British Journal of Educational Technology*, vol. 42(1), pp. 40–49.

Agostino, D. & Sidorova, Y. (2016). A performance measurement system to quantify the contribution of social media: new requirements for metrics and methods. *Measuring Business Excellence*, vol. 20(2), pp. 38–51.

Akshay. (2021). *Automate Web Scraping Using Python AutoScraper Library*. *Analytics Vidya* [online]. [Accessed 23 April 2021]. Available at: https://www.analyticsvidhya.com/blog/2021/04/automate-web-scraping-using-python-autoscraper-library/.

Alhawiti, K. M. (2014). Natural Language Processing and its Use in Education. *International Journal of Advanced Computer Science and Applications*, vol. 5(12), pp. 72–76.

Anglin, K. L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. *Journal of Research on Educational Effectiveness*. Routledge, vol. 12(4), pp. 685–706.

Ayyappan, S. & Matilda, S. (2020). Criminals and Missing Children Identification Using Face Recognition and Web Scrapping. *2020 International Conference on System, Computation, Automation and Networking, ICSCAN 2020*, pp. 1–5.

Bansal, S. (2017). *Ultimate Guide to Understand and Implement Natural Language Processing*. *Analytics Vidya* [online]. [Accessed 15 April 2021]. Available at: https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G. & Viviani, M.

(2018). WoLMIS: a labor market intelligence system for classifying web job vacancies. *Journal of Intelligent Information Systems*. Journal of Intelligent Information Systems, vol. 51(3), pp. 477–502.

Capiluppi, A., Serebrenik, A. & Singer, L. (2013). Assessing technical candidates on the social web. *IEEE Software*, vol. 30(1), pp. 45–51.

Chen, Y. (2010). Natural Language Processing in Web data mining. *Proceedings - 2010 IEEE 2nd Symposium on Web Society, SWS 2010*. IEEE, pp. 388–391.

D'Haen, J., Van Den Poel, D., Thorleuchter, D. & Benoit, D. F. (2016). Integrating expert knowledge and multilingual web crawling data in a lead qualification system. *Decision Support Systems*. Elsevier B.V., vol. 82, pp. 69–78.

Dadzie, A. S., Sibarani, E. M., Novalija, I. & Scerri, S. (2018). Structuring visual exploratory analysis of skill demand. *Journal of Web Semantics*. Elsevier B.V., vol. 49, pp. 51–70.

Dhanith, P. R. J., Surendiran, B. & Raja, S. P. (2020). A Word Embedding Based Approach for Focused Web Crawling Using the Recurrent Neural Network. *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. In Press(In Press), p. 1.

Diouf, R., Sarr, E. N., Sall, O., Birregah, B., Bousso, M. & Mbaye, S. N. (2019). Web Scraping: State-of-the-Art and Areas of Application. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pp. 6040–6042.

Dogucu, M. & Çetinkaya-Rundel, M. (2020). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics Education*. Taylor & Francis, vol. 0(0), pp. 1–24.

Drake, M. S., Thornock, J. R. & Twedt, B. J. (2017). The internet as an information intermediary. *Review of Accounting Studies*. Review of Accounting Studies, vol. 22(2), pp. 543–576.

Esposito, M., ElSholkamy, M. & Fischbach, T. (2017). *The UAE and The Future of Work. First look.*

Frazier, K., Davis, H. & Vickery, J. (2020). Seeing the Forest for Trees: Tools for Analyzing Faculty Research Output. *Serials Review*. Routledge, vol. 46(3), pp. 184–189.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, vol. 15(5), pp. 788–797.

Gupta, K., Mittal, V., Bishnoi, B., Maheshwari, S. & Patel, D. (2016). AcT: Accuracy-aware crawling techniques for cloud-crawler. *World Wide Web*, vol. 19(1), pp. 69–88.

Han, S. & Anderson, C. K. (2021). Web Scraping for Hospitality Research: Overview, Opportunities, and Implications. *Cornell Hospitality Quarterly*, vol. 62(1), pp. 89–104.

Hillen, J. (2019). Web scraping for food price research. *British Food Journal*, vol. 121(12), pp. 3350–3361.

Hong, S. H., Lee, S. K. & Yu, J. H. (2019). Automated management of green building material information using web crawling and ontology. *Automation in Construction*. Elsevier, vol. 102(April 2018), pp. 230–244.

Isahara, H. (2007). *Resource-based Natural Language Processing*. *Computational Linguistics* [online].Available at: http://ieeexplore.ieee.org/document/4368002/.

Johnson, M. E., Albizri, A. & Jain, R. (2020). Exploratory Analysis to Identify Concepts, Skills, Knowledge, and Tools to Educate Business Analytics Practitioners. *Decision Sciences Journal of Innovative Education*, vol. 18(1), pp. 90–118.

Jorge, O., Pons, A., Rius, J., Vintró, C., Mateo, J. & Vilaplana, J. (2020). Increasing online shop revenues with web scraping: a case study for the wine sector. *British Food Journal*, vol. 122(11), pp. 3383–3401.

KHDA. (2018). Know How QAD ensures the Quality of Qualifications. Knowledge and Human Reseource Authority [online].Available at: https://www.khda.gov.ae/CMS/WebParts/TextEditor/Documents/Know_How_QAD_ensures_the_Quality_of_Qualifications_QAD_Book_06_Eng.pdf.

Krotov, V. & Tennyson, M. (2018). Research Note: Scraping financial data from the web using the R language. *Journal of Emerging Technologies in Accounting*, vol. 15(1), pp. 169–181.

kumar, R., Jain, A. & Agrawal, C. (2016). Survey of Web Crawling Algorithms. *Advances in Vision Computing: An International Journal*, vol. 3(3), pp. 1–7.

Li, W., Wang, S. & Bhatia, V. (2016). PolarHub: A large-scale web crawling engine for OGC service discovery in cyberinfrastructure. *Computers, Environment and Urban Systems*. Elsevier Ltd, vol. 59, pp. 195–207.

Litecky, C., Aken, A., Ahmad, A. & Nelson, H. J. (2008). Mining for Computing Jobs. *IEEE Software*, pp. 78–85.

Lloyd, S. (2008). The digital universe. *Physics World*, vol. 21(11), pp. 30–36.

Lo, H. C., Koedijk, K. G., Gao, X. & Hsu, Y. T. (2020). How do job vacancy rates predict firm performance? A web crawling massive data perspective. *Pacific Basin Finance Journal*, vol. 62(2271), pp. 1–43.

Lovaglio, P. G., Cesarini, M., Mercorio, F. & Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 11(2), pp. 78–91.

Lovaglio, P. G., Mezzanzanica, M. & Colombo, E. (2020). Comparing time series characteristics of official and web job vacancy data. *Quality and Quantity*. Springer Netherlands, vol. 54(1), pp. 85–98.

Macapinlac, T. (2019). The Legality of Web Scraping: A Proposal. *Federal Communications Law Journal*, vol. 71(3), pp. 399–0_7,0_8 [online].Available at: https://search.proquest.com/scholarly-journals/legality-web-scraping-proposal/docview/2326822684/se-2?accountid=14511%0Ahttps://ucl-new-primo.hosted.exlibrisgroup.com/openurl/UCL/UCL_VU2?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=a.

Massimino, B. (2016). Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data. *Journal of Business Logistics*, vol. 37(1), pp. 34–42.

Mauro, A. De, Greco, M., Grimaldi, M. & Ritala, P. (2018). Human resources for Big Data professions : A systematic classification of job roles and required skill sets. *Information Processing and Management*. Elsevier Ltd, vol. 54(5), pp. 807–817.

Maurya, B. B. P., Ray, A., Upadhyay, A., Gour, B. & Khan, A. U. (2019). Recursive Stock Price Prediction with Machine Learning and Web Scrapping for Specified Time Period. *IFIP International Conference on Wireless and Optical Communications Networks, WOCN*, pp. 55–57.

Menke, A. & Giehl, K. (2021). Collecting data on textiles from the internet using web crawling and web scraping tools. *Forensic Science International*. Elsevier, p. 110753.

Miranda, C. (2020). *4 Ways Web Scraping Can Help Your Business*. *ParseHub Blog* [online]. [Accessed 4 April 2021]. Available at: https://www.parsehub.com/blog/web-scraping-help-business/.

Nashipudi, P. (2020). Developing Framework of Web Scraper for Agriculture Data using Client Server Module. *International Journal of Computer Engineering In Research Trends (IJCERT),ISSN*, (8), pp. 2349–7084.

NLP - Linguistic Resources. (2021) [online]. [Accessed 15 April 2021]. Available at: https://www.tutorialspoint.com/natural_language_processing/natural_language_processing_linguistic_resources.htm.

NQA. (2012). *Qualifications Framework for the Emirates Handbook* [online].Available at: www.nqa.gov.ae.

NQA. (2014). *VETAC Q + NOSS SYSTEM GUIDELINES VETAC ' s Quality Assurance and Endorsement System to develop and deliver National Qualifications*.

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*. IEEE, vol. 7, pp. 157595–157613.

Pazmiño, A. M. & Diaz, C. O. (2018). Analysis and Data Visualization of Labor Demand in Ecuadorian Territory. *2018 ICAI Workshops (ICAIW) IEEE*, pp. 1–7.

Perez, M. (2019). *Is Web Scraping Legal? ParseHub Blog* [online]. [Accessed 2 April 2021]. Available at: https://www.parsehub.com/blog/web-scraping-legal/.

Phaphuangwittayakul, A., Saranwong, S., Panyakaew, S. N., Inkeaw, P. & Chaijaruwanich, J. (2018). Analysis of Skill Demand in Thai Labor Market from Online Jobs Recruitments Websites. *Proceeding of 2018 15th International Joint Conference on Computer Science and Software Engineering, JCSSE 2018*. IEEE, pp. 1–5.

Pillai, P. & Amin, D. (2020). Understanding the requirements of the Indian IT industry using web scrapping. *Procedia Computer Science*. Elsevier B.V., vol. 172, pp. 308–313.

Puckett, J., Davidson, J. & Lee, E. (2012). Vocational education: The missing link in economic development [online].Available at: https://www.bcgperspectives.com/content/articles/education_public_sector_vocational_education/

.

Rahman, A. & Al, J. (2016). An Overview - Indicators of the Vocational Education Sector in UAE. *International Journal of Scientific & Engineering Research*, vol. 7(6), pp. 995–1001.

Salem, M. & Alawi, A. (2016). Culture and Impact of Culture in Operations Management in Vocational Education Sector of United Arab Emirates. *International Journal of Scientific Engineering and Research*, vol. 4(11), pp. 67–70.

Selvadurai, J. (2013). A Natural Language Processing based Web Mining System for Social Media Analysis. *International Journal of Scientific and Research Publications,* vol. 3(1), pp. 2250–3153 [online].Available at: www.ijsrp.org.

Single, J. I., Schmidt, J. & Denecke, J. (2020). Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing. *Safety Science*. Elsevier, vol. 129(May), p. 104747.

Singrodia, V., Mitra, A. & Paul, S. (2019). A Review on Web Scrapping and its Applications. *2019 International Conference on Computer Communication and Informatics, ICCCI 2019*. IEEE, pp. 1–6.

Sundaramoorthy, K., Durga, R. & Nagadarshini, S. (2017). NewsOne - An Aggregation System for News Using Web Scraping Method. *Proceedings - 2017 International Conference on Technical Advancements in Computers and Communication, ICTACC 2017*, pp. 136–140.

Surabhi, M. C. (2013). Natural language processing future. *2013 International Conference on Optical Imaging Sensor and Security, ICOSS 2013*. IEEE, pp. 7–9.

Szabó, I. (2011). Comparing the competence contents of demand and supply sides on the labour market. *33rd International Conference on Information Technology Interfaces*, pp. 345–350.

Tatwadarshipn. (2021). *Role of Machine Learning in Natural Language Processing*. *Analytics*

Vidya [online]. [Accessed 23 April 2021]. Available at: https://www.analyticsvidhya.com/blog/2021/04/role-of-machine-learning-in-natural-language-processing/.

Thi, P. Q., Diep, H. T., Thao, N. D., Pham-Nguyen, C., Dinh, T. Le & Nam, L. N. H. (2020). Towards An Ontology-Based Knowledge Base for Job Postings. *Proceedings - 2020 7th NAFOSTED Conference on Information and Computer Science, NICS 2020*, pp. 267–272.

Ulbricht, L. (2020). Scraping the demos. Digitalization, web scraping and the democratic project. *Democratization*. Taylor & Francis, vol. 27(3), pp. 426–442.

Wikipedia. (2018). *Social networking service* [online]. [Accessed 2 April 2020]. Available at: https://en.wikipedia.org/wiki/Social_networking_service.

Yang, S., Wi, S., Park, J. H., Cho, H. M. & Kim, S. (2020). Framework for developing a building material property database using web crawling to improve the applicability of energy simulation tools. *Renewable and Sustainable Energy Reviews*. Elsevier Ltd, vol. 121(January), p. 109665.

Yash. (2020). *A Quick Guide to Text Cleaning Using the nltk Library*. *Analytics Vidya* [online]. [Accessed 23 April 2021]. Available at: https://www.analyticsvidhya.com/blog/2020/11/text-cleaning-nltk-library/.