Addressing the Gap Between the University's Curriculum Related to Data Science Degree and Industry Needs

Project Proposal

Contents

Title : Addressing the gap between the university's curriculum related to data science degree and industry needs

Abstract :

Any job market is seeking best possible talents from the educational institutes all over the world. However, there is a concern in society that the demand for the job market can not be fulfilled by the educational institutes due to the lack of proper standard of the university curriculums towards the job demand. And, data science has been an emerging area recently which is looking for best possible skills to cater the demand. Therefore, there should be a specific method to upgrade universities curriculums based on the dynamic change of the job market of the Data science. The research project will answer that question by implementing a proper recommendation system based on the machine learning and natural language processing. Text clustering, text classification, machine leaning algorithm like KNN algorithm and ranking based algorithms will be used in this research. Finally, the results of this thesis will be applied to enhance the university curriculum in order to fulfill the demand of the job market of the data science with fact-based decision making.

Key word : Education, Job Demand, Machine learning, Natural language processing

Introduction :

The purpose of the education is to facilitate a process to get learning, knowledge and skill. Basically, In the nutshell, the education system should have all the capabilities to meet the requirement of the job market. If not, it can be identified as a waste. On the other hand, education can be considered as the society's cornerstones. And, education creates useful and productive components of the societies. However, when market of the jobs shifts in dynamically, the education should change dynamically to meet to the demand of the dynamic job market. (Griffin, 2016). Measuring the speed of job dynamics (Skills and Job nature) and the speed of the curriculum changes in the education degrees to meet the job market requirement is a difficult task with out proper technological layer. The technologies can be implemented in this area to monitor the dynamic nature of the job market. Specially, Natural language processing (NLP) technologies with cloud environments will be a north worthy component to solve this problem (Wang, 2017). According to the latest literature, most of the researches have not done research in these areas related to the demand supply matching in between industry need and curriculum related to the university. The real gap in this field have been created because, ranking the universities curriculum to the need of the job market is a difficult task and special machine learning based recommendation system could be solved this problem. Based on this research, this will mainly fulfill this gap with the latest technologies of machine leaning (ML) based in NLP.

Problem Statement and Aim :

There is a recent surge of demand for data scientists in the industry and the response from the education institutions in terms of contents because the content in a data science degree is relatively slow. This gap will fulfill with the latest technological application of machine learning and NLP based on this research therefore that would be the main research aim of this research project.

Project Objectives :

Use ML and NLP to analyze the gap between the industry need (by analyzing job ads) and university curriculums (by analyzing teaching materials),

- Identify the tools and techniques to scrape the demand (Industry need) and supply information (university curriculums) and matching them based on the ranking and summarizations.
- Design and create new courses and programs with the insights generated from scraped data.

Literature Review :

Text Mining and NLP in the field of Data Science

The things that are expressed in the everyday carries huge amount of information. These information can be derived from both verbally and in written format. The topics that are spoken, the tone, comments, statements of words are specially related to the amount of information. However, they are in their own format of unstructured manner. That information doesn't have proper structure and data frames like the numerical data and information (Yse, 2019). The Unstructured data can be clearly distinguished from the traditional row and column structure of data storing. In other words, it is significantly messy and need some extra effort to manipulate data to generate more insights(Marr, 2019).

In the simple terms, the natural language like text or speech can be automatically handled and that process is called as Natural Language Processing (NPL). The technological layer has been done a north worthy attempt on the NPL to derive most prominent insights on top of the use cases (Yse, 2019). The following are some use cases related to the NPL. The most common approach is Machine Translator for the languages. Specially google translator used machine learning approaches to deal with language translator from one language to another. Secondly, the speech translator, sentiment analysis for labelling the review to check whether that review is positive or negative, chatbots and Automatic summarizations can be identified as the most prominent approaches in the field of Text Mining and NLP (Kharkovyna, n.d.)

Document Clustering : Summarizations

The group of documents can be summarized using these techniques into various clusters based on their similar content. The main idea of that is the descriptors which have been built using set of words that explains a common content within the cluster. This unsupervised method will focus on the classify the objects into several groups and that groups can be called as the clusters. In simple terms, dividing the similar text into same group or cluster is the main idea of this techniques (Anon., 2020). In this research, the clustering will be performed on the job advertisement or degree programs, it makes sure that all the matching similar documents are combined into one cluster or group. Because, in the web scraping, most of the time, large amount of data will be scraped and in order to summarize the documents, the summarization application will be used. (Arthur, 2020) has used the test summarization techniques in the research project on software development source code documentation summarizations.

Text Categorizations

This is a special area in the NLP, because, this method uses predefined categories for documentations and some reviews. This method will help to improve the decision making because, predefined text classification/text categorization will direct end users to predict the future text field or text review on their product or services (Anon., 2020). In this research project, several reviews of degree programme should be scraped. Once the reviews have been scraped from the several sources, they will be stored in a big data platform to decision making. When comes to the decision making, each and every review should identify as a positive review or negative review. Because, reading all the large amount of review is not a practical scenario and time-consuming process(Tao, et al., 2020).

Past researches

When focusing on the similar research,(Maldonado & Seehusen, 2018) has done a research project to analysis the student choices for certain degree programme. In this research project, a feedback generation mechanism has been tested with students who have completed the degree. The techniques wise that research project has been used clustering approach for reviews from the student and identified some significance clustering and those insights have been used to improve the degree curriculum in the university as a decision making. (Pejic-Bach, et al., 2020) has done a research project based on the Text mining on the Industry 4.0 job applications .This method will be directly applicable for identify the new immerging talents in organizations to build an excellent work force which come to the matching for future demand (Pejic-Bach, et al., 2020). The job classification is a significant area in the field of NLP and new algorithms and novel techniques in this field will help to get clear idea on market dynamic in the job market (Boselli , et al., 2018) and this project has used text classification using machine learning. These approaches will support to take smart decisions clearly.

Technologies that can be used

(Raman, 2019) has done an implementation to screen the Data science resumes. Basically, a specific dictionary with skill sets of categories related to the data science has been set up then specific NLP algorithm has been used to search on top of the resumes for the list of words that have been mentioned in the skill set of categories. Based on that, the algorithm will be providing the count of occurrences of the words in the list. Further to that (Roy, et al., 2020) has done a research project on resume recommendation based on machine learning approaches to select the best candidates. All the resumes will classify into different categories using specific classifications. Once the resumes have been classified into different clusters, a ranking approach will be applied to rank the candidates in each category. K-nearest neighbors' algorithm (K-NN) based machine learning approach has been used to recommend the best suitable candidates to each job classifications. According to the (Guo, 2016), job matching system has been completed fulfilling a gap to match suitable jobs for a candidate. Multiple amount of jobs is published in the job advertising portals and ranking the suitable jobs based on the candidate's desire and qualification is easy task based on this application with novel technologies.

Researches on web scrapings

(Dewi, et al., 2019) has introduced a novel approach to enhance the web scraping capabilities to retrieve more information. According to the (Dewi, et al., 2019), Information redundancy of the information is the main barrier of the web scarping applications and the method proposed by this method will enhance the search capabilities, combining the information in better way and presenting with better approach.

Methodology :

Web scraping

This has been considered as a method to extract the information which have been in a website. And, this scraping activity can be done with or without agreement of the owner of the web sites.

The web scrapping can be identified as two methods mainly. The manual and automated methods. Basically, manual scrapping involves the method-based copy paste the information however, according to the past experiences of the several literatures, this has been identified as a time-consuming method. In that mind set, this research project will be used the following automated methods as the web scraping methods.

Google Sheets: This can be used as a scraping tool and IMPORTXML() function can be used scrape the data from the website. This method can be mainly using for the situation where the end user wants specific pattern or data to be scraped from a website.

DOM Parsing: This is referred to the Document Object Model or DOM and the structure of the XML files are defined based on the structure, style and content of the XML files. If scrapers want the details structure of the web page, the DOM parser will be used. DOM Parser is used to get the nodes containing information and XPath can be directly used to scrape the web pages.

HTML Parsing: This is mainly based on the JavaScript and Nested HTML Pages. Resource extractions from the web, Link extractions, screen scraping a text extraction are mainly done based on this method.

Document or text clustering with K- mean/ KNN algorithm

Natural Language Toolkit based on the python can be directly used for this purpose. NLTK has many features to do the data preprocessing tasks like Stop words, Stemming, and Tokenizing and once the job description or university curriculums have been scraped these text prepressing tasks should be applied to improve the algorithms performances tasks related to the machine learning clustering. Then, frequency-inverse document frequency (tf-idf) will eb created. After that it converts the synopses list into a tf-idf matrix. Once all those proses have been done, the algorithms like K- mean/ KNN will be applied. Those have been identified as the clustering algorithms based on the machine learnings. Python based sklearn package has all the capabilities to performance these tasks. Hierarchical document clustering is another approach to deal with this situation. Based on the above approaches, the text documents like the job description or university curriculums will be group into the similar clusters. This will be the main technique in this research project to come up the comprehensive insights.

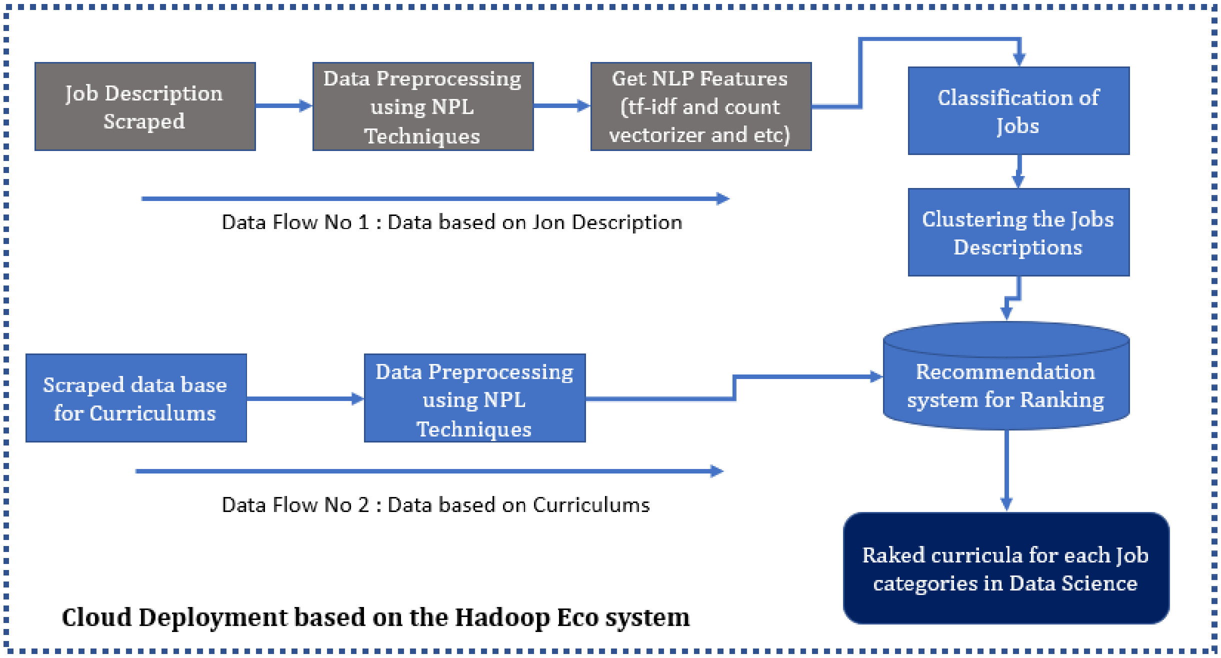
Techniques than can be used to screening university curriculum with the Job market demand

(Raman, 2019) has mainly used specific technique to screen candidate's Resume and for this research also that can be used to screen the university curriculum and the job market descriptions of the certain jobs. First, all the jobs advertisements related to the Data science should be scraped and clustering application can be applied on top of the scraped advertisements. Because, Data science in nowadays has many areas like Machine learning and AI development, Data Evangelists, Data Engineers, Data Analysts etc. Once the document clustering has been done on the scraped text related to the job advertisements. Those data science fields can be identified . Once the above process has been done, the following approaches can be used. Based on the NLP techniques on text preprocessing a dictionary or table which has all the various skills sets categorized for various jobs clusters of data science can be created. Then a specific NLP algorithm that parses whole curriculum of universities and generally screen for the words that have been mentioned in the dictionary. The results would be the summary results based on the above screening process. Then the results can be used to get decisions about matching capabilities of each university curriculum for specific job cluster of the Data science field as the results.

The matching process of the curriculum and the job description can be done with above process however, further to that, non parametric algorithms in machine learning for classifications (Support Vector machine, Decision Tree, Gradient Boosting and Neural networks) can be used to classifies each job advertisement in to good job or bad job based on the NLP techniques for classifications. Then, based on the descriptions of the job advertisements, the top match university curriculum could be ranked with the help of Content-based recommendations. Basically, the k-NN algorithm and cosine similarity can be applied to find the nearest curriculum (Most matching) to a specific job description as the results of ranking. Based on the results of curriculum ranking for each and every job clusters in data science, the end user can be identified the contribution from each university curriculum to fulfill the demand of the job market.

The following is the high-level diagram for this research project. The high-level diagram has two data flows, and this will be set up in the cloud big data analytics platform to easily deploy the applications as whole product. The job descriptions / Job market scraped documents will be coming in one direction and curriculum related scraped documents will be coming in another directions. Once the preprocessing tasks have been done, specific mathematical recommendation system will be created based on the ranking related algorithms.

High Level Conceptual Design


Job Description Scraped
Data Preprocessing using NPL Techniques
Get NLP Features (tf-idf and count vectorizer and etc)
Classification of Jobs
Clustering the Jobs Descriptions
Data Flow No 1 : Data based on Jon Description
Scraped data base for Curriculums
Data Preprocessing using NPL Techniques
Recommendation system for Ranking
Data Flow No 2 : Data based on Curriculums
Raked curricula for each Job categories in Data Science
Cloud Deployment based on the Hadoop Eco system

Figurer 1.1

Conclusion and summary :

In order to full fill the research aim, The scrapped texts related to the job description and university curriculums will be summarized based on the clustering/ classification approaches in NPL based machine learning and then the process of matching the job demand to the university curriculum will be done based on the recommendation approaches in machine learning and final product will be deployed in the cloud environment for easy deployment. As the results of this applications, educational institutes can enhance their curriculums of data science based on the current and future demands.

References :

Arthur, M. P., 2020. Automatic Source Code Documentation using Code Summarization. Procedia Computer Science , Volume 171, pp. 2522-2531.

Guo, S., 2016. RESUMATCHER: A PERSONALIZED R ´ ESUM ´ E-JOB MATCHING SYSTEM. Expert Systems with Applications, Volume 60.

Raman, V., 2019. How I used NLP (Spacy) to screen Data Science Resume. [Online] Available at: https://towardsdatascience.com/

Anon., 2020. Text mining methods and techniques: The beginner’s guide, s.l.: https://roboticsbiz.com/.

Boselli , R., Cesarini, M., Merco, F. & Mezzanzanica, M., 2018. Classifying online Job Advertisements through Machine Learning. Future Generation Computer Systems, Volume 86, pp. 319-328.

Dewi, L. C., M. & Chandraa, A., 2019. Social Media Web Scraping using Social Media Developers API and Regex. Procedia Computer Science, Volume 157, pp. 44-449.

Griffin, M., 2016. https://www.cio.com/. [Online] Available at: https://www.cio.com/article/3095287/the-future-of-jobs-and-education.html

Kharkovyna, O., n.d. Natural Language Processing (NLP): Top 10 Applications to Know, s.l.: https://towardsdatascience.com/.

Maldonado, E. & Seehusen, V., 2018. Data mining student choices: A new approach to business curriculum planning. Journal of Education for Business, 93(5), pp. 198-203.

Marr, B., 2019. What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone, s.l.: https://www.forbes.com/.

Pejic-Bach, M., Bertoncel, T. & Meško, M., 2020. Text mining of industry 4.0 job advertisements. International Journal of Information Management, Volume 50, pp. 416-431.

Roy, P. K., Chowdhary, S. S. & Bhatiab, R., 2020. A Machine Learning approach for automation of Resume Recommendation system. Procedia Computer Science, Volume 167, pp. 2318-2327.

Tao, C., Guo, H. & Huang, Z., 2020. Identifying security issues for mobile applications based on user review summarization. Information and Software Technology, Volume 122.

Wang, K., 2017. Text mining technology based on cloud computing. Acta Technica CSAV (Ceskoslovensk Akademie Ved), Volume 62, pp. 72-83.

Yse, D. L., 2019. Your Guide to Natural Language Processing (NLP), s.l.: https://towardsdatascience.com/.