JOB RECOMMENDATION SYSTEM

Ananya Pandey - 30757924

FIT5145: DATA SCIENCE PROJECT REPORT

TABLE OF CONTENTS

PROJECT DESCRIPTION:

The great picture of a job searcher poring over a nearby newspaper's job listings is currently a thing of the past. Recently, job seeking and enlisting sites have been confronting a striking ascent. According to records, almost 50% of ongoing recruits (41%) discovered their present job through an online activity board. (Delgado, 2019)
One such frontrunner in Australia, *Seek Limited* and its subsidiary companies, known as the *Seek Group*, center around encouraging the coordination among jobseekers and business openings and helping hirers discover appropriate candidates for the promoted jobs. *SEEK* works for a beneficial outcome on a worldwide scale, with an introduction to 2.9 billion individuals, over 51 million students, and a nearness in 18 nations including China and across South-East Asia and Latin America ("About *SEEK*", 2019). *SEEK*, with its innovative business model, has provided a platform based on more than 22 years of commercial experience, along with a combination of emerging fields like Artificial Intelligence and Machine Learning. *SEEK* gets more than 41 million visits every day by potential job seekers and around 900,000 jobs are posted by various companies looking for employers to join their firms ("About *SEEK*", 2019). As the measure of data grows, an improved suggestion-framework with advanced features is important to help match the right candidate with the right job. Such platforms do provide a list of jobs according to the users' preference but the list does not always cater to the needs of the user and sometimes the user himself is not very clear of what positions he can apply for according to his achievements. The resumes are not fully utilized when a user uploads their resume on their portal. Enhanced and improved recommendations can help even the most modern client find extra occupations that their quests would not have revealed.

Data Science Roles:

1. Data Scientist: A Data Scientist at *SEEK* is an individual having enthusiasm for tackling genuine issues and ability in Python/R and Apache Spark. In general, the job of a Data Scientist is to address the whole procedure to get important bits of knowledge.
2. Data Analyst: As a Data Analyst, *SEEK* anticipates that the individual should design project roadmaps and plan data integration activities and have solid marketing prudence. The individual ought to be knowledgeable about heading data-driven groups with different technical strengths.
3. Data Engineer: The primary assignment of data engineers is to give a dependable framework for data and have the ability in the field of software engineering and backend development. The data engineer is responsible for the data mining process as well.
4. Data Architect: For *SEEK*, a Data Architect ought to be a certified individual having experience in this field which requires him/her to make, plan, and maintain the architecture of *SEEK*.
5. Statistician: As a statistician at *SEEK*, an individual should have a strong background and working experience in the same field, and should be able to do statistical analysis by applying probabilistic models for the system to be designed.

BUSINESS MODEL:

By employing the modern data science approach, this project aims at building an improvised version of the currently being used recommendation system at *SEEK* which will be valuable to the job seekers as well as the hiring company, thus providing a positive impact to the society.

Specifically, job recommender systems are intended to retrieve a list of occupation positions to a job applicant depending on his/her preferences. Due to existing systems data handling techniques, effective recommendation results are not up to the mark. To acquire better recommendation results, a few new approaches are presented to overcome the concerns highlighted in the report. This recommendation system retrieves a list of job positions applicable to a job applicant by matching the job description presented by the hirer to the profile of the user which can be determined from the user's resume i.e. classifying the domain in which the user can work in according to their skills and qualifications. It will also list out the job positions available according to the location of the user which can be found from the IP Address that the user is using for logging in to their *SEEK* portal. If a user does not get selected for a particular job that he had applied to, the system compares the resume of the rejected applicant to that of the accepted one to present the distinguishing features between the accepted and the rejected profiles which can be extremely beneficial for people looking for jobs.

Challenges:

- Every day millions of job hirers post new jobs. This implies that every time new job posts are added, the set of recommendable jobs keeps on changing.
- As millions of new job seekers visit *SEEK* every day and begin their job search, the recommendation system should be able to generate recommendations with very limited user data.
- The job that gets posted on *SEEK* should be timely. A job that has expired with respect to its employment status should be inaccessible and not be recommended to an applicant.
- Ensure each job posting receives a sufficient amount of applications to assure a recruitment.

Benefits:

- This recommendation system can bring in a lot more people to the website, thus increasing the traffic to the *SEEK*.
- Job seekers will end up being more engaged as they get better individualized job recommendations, thereby assisting in picking correct job options for the user.
- This system will also help the seeker to discover the domains where he/she can work.
- It is extremely easy for the recruiter to put up a job description and get the candidates with the maximum matching percentage for the requirements.

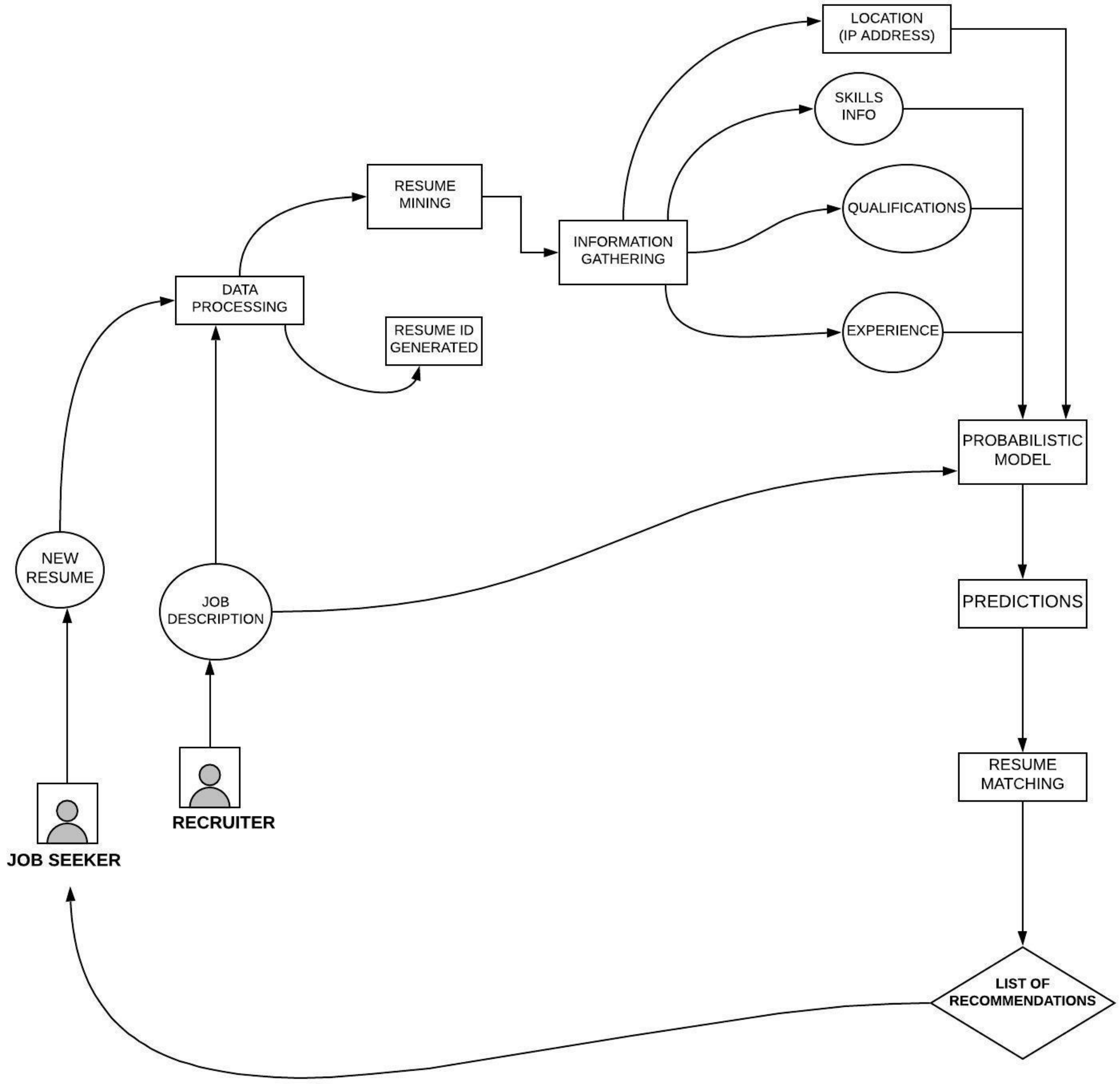


Figure 1: Influence diagram showing the basic model of the system.

CHARACTERIZING THE DATA AND DATA PROCESSING:

The data can be characterized based on the following features -

1. **Raw data**: Data which is actually recorded through the resume that each user submits while signing up into their SEEK portal.
2. **Meta-data**: It includes the data description i.e. the author of the document, the person who last modified it, the editing time, any comments etc.

Data Characteristics	Description
Data Source	SEEK gets its data from the various job applications, the job seekers' profile, the hirer's requirements. Data about the jobs that exist and the skills needed for each is also collected.
Data Type	The data is mostly textual. It can come in various formats i.e. a word document, or a pdf file.
Data Volume	SEEK gets at least 15 million visits per month. It is difficult to measure the actual size of the data as much information regarding this is not provided by SEEK.com. This makes SEEK have a huge amount of data at present which includes the original data of users in the form of resumes, the position requirements and their backups as well.
Data Variety	Data has a lot of variety as each person is different with different backgrounds and experiences. In a similar way, new job positions are created every day which shows the variety in the jobs posted as well.
Data Velocity	As there are thousands of people, looking for job at the same time, for the purpose of predictive modelling, this job recommender system will provide real – time analysis as well as provide weekly recommendations to the user. The data input has to be updated continuously to be able to predict correctly.
Data Veracity	It is the authenticity of the data i.e. whether the job posting is from a credible source and is not a scam or a fake company, and whether the profiles created are not fake, by verifying them through their LinkedIn or GitHub profiles.

Data Processing:

Being a data science project, data processing (which includes pre-processing) plays an important role as it is the foundation step for providing any meaningful information from the given data at any time. Once the user registers and uploads their resume, the collected data is used for the mining process.

Data at SEEK is extremely challenging because it is increasing every moment due to the resumes being uploaded by new users. The processing stage isn't just limited to require the input from the user and transform it into a structure which is suitable for analysis and presenting the predictions accordingly. The complex nature of the tasks makes Apache Spark the ideal choice for this project to process the big amount of information in an efficient manner. Spark places the info within the memory to be analyzed, giving it significant speed. Spark can connect with big databases, the HDFS or perhaps cloud based storage.

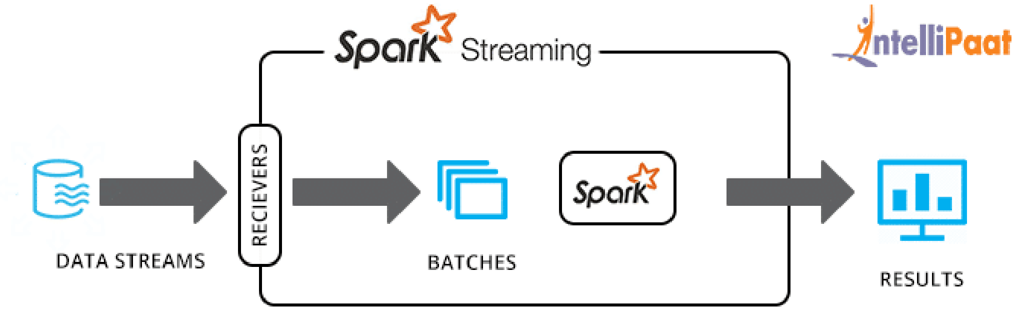


Figure 2: Working of Apache Spark
("A Guide to Apache Spark Streaming - Intellipaat Blog", 2017).

The data resources that have been collected will be further integrated and cleaned.

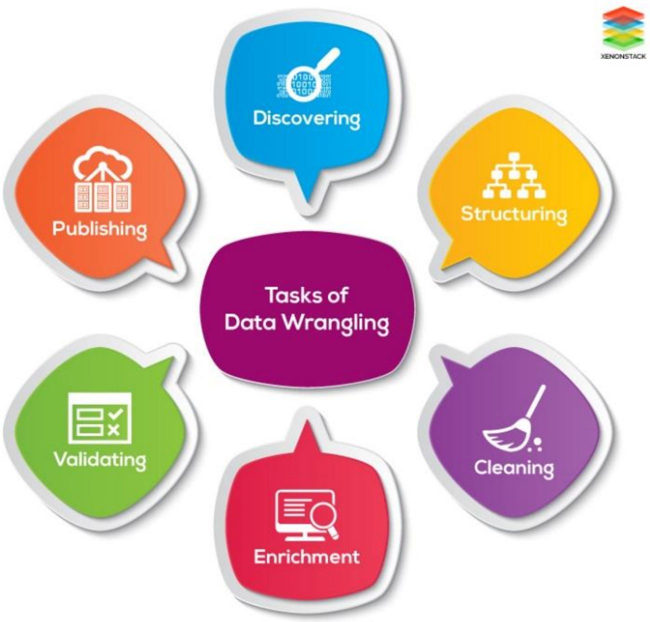


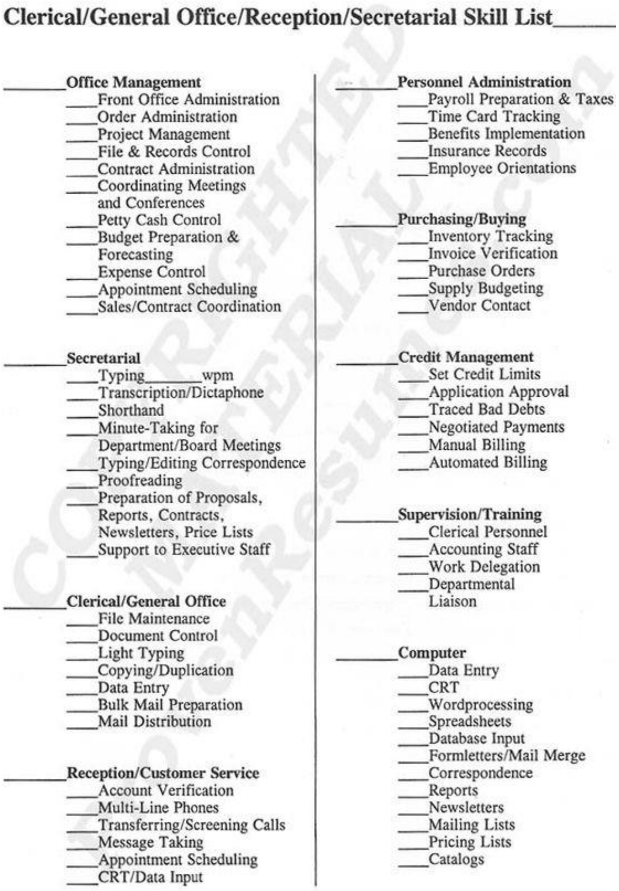
Figure 3: Data Wrangling Process.
(Gill, 2018)

During the process of data wrangling, the input data is first converted in a same text format as different files can be in a different format and then cleaned in a manner, that it does not contain many null values, remove extra spaces, look for the words that will be needed for resume matching, there are no grammatical errors which can be a problem later on. The data is structured according to the various attributes in the resume i.e. the education background, the qualifications, the skills etc. This can be done using the ‘tm’ package from the R programming language, which is suitable for text mining.

RESOURCES:

For this project, the significant potential resource will be the data obtained through resumes and job postings at SEEK. The data for various types of jobs available anywhere is also collected to know the types of positions that are known and are available (skills, 2020).

A sample data for this data science project has been shown below:



Clerical/General Office/Reception/Secretarial Skill List_____

_______**Office Management**
____Front Office Administration
____Order Administration
____Project Management
____File & Records Control
____Contract Administration
____Coordinating Meetings and Conferences
____Petty Cash Control
____Budget Preparation & Forecasting
____Expense Control
____Appointment Scheduling
____Sales/Contract Coordination

_______**Secretarial**
____Typing_______wpm
____Transcription/Dictaphone
____Shorthand
____Minute-Taking for Department/Board Meetings
____Typing/Editing Correspondence
____Proofreading
____Preparation of Proposals, Reports, Contracts, Newsletters, Price Lists
____Support to Executive Staff

_______**Clerical/General Office**
____File Maintenance
____Document Control
____Light Typing
____Copying/Duplication
____Data Entry
____Bulk Mail Preparation
____Mail Distribution

_______**Reception/Customer Service**
____Account Verification
____Multi-Line Phones
____Transferring/Screening Calls
____Message Taking
____Appointment Scheduling
____CRT/Data Input

_______**Personnel Administration**
____Payroll Preparation & Taxes
____Time Card Tracking
____Benefits Implementation
____Insurance Records
____Employee Orientations

_______**Purchasing/Buying**
____Inventory Tracking
____Invoice Verification
____Purchase Orders
____Supply Budgeting
____Vendor Contact

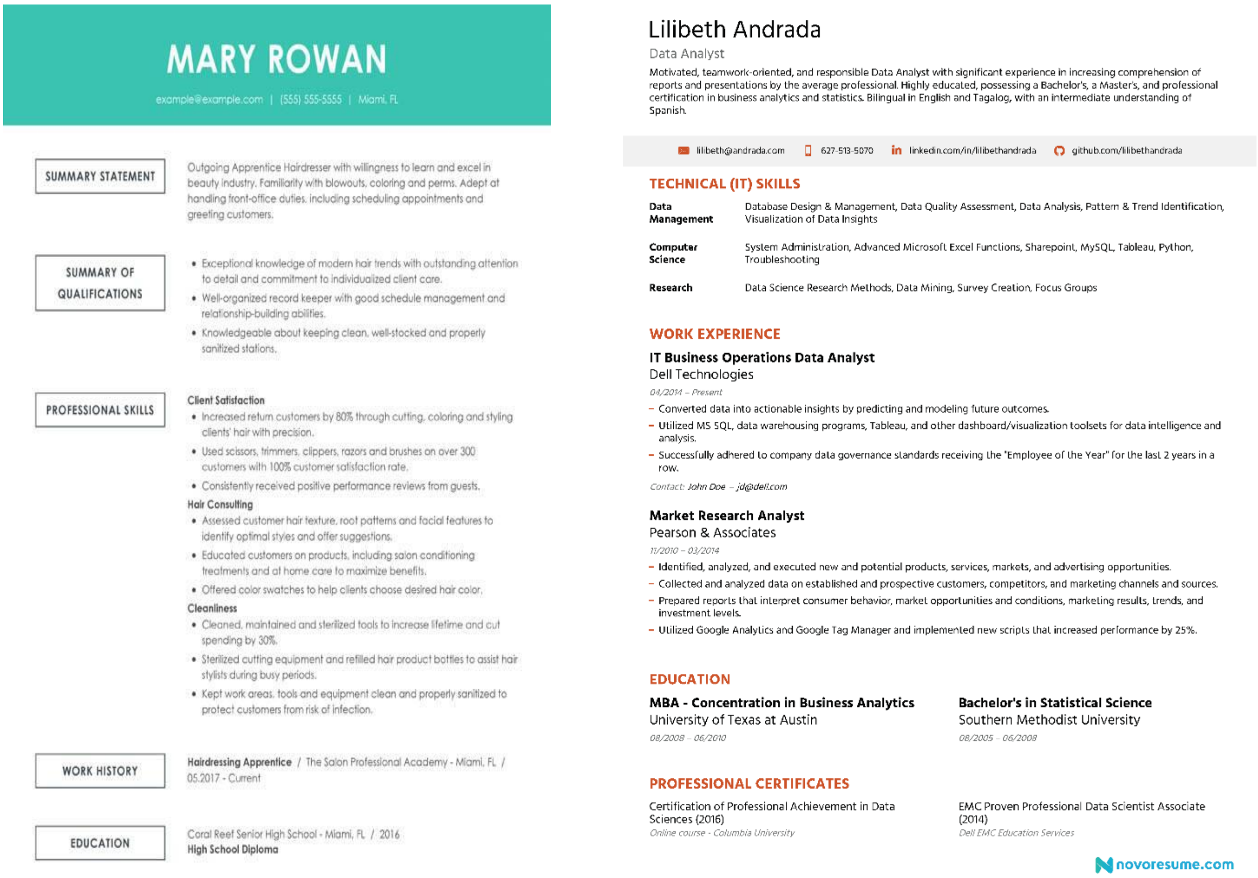
_______**Credit Management**
____Set Credit Limits
____Application Approval
____Traced Bad Debts
____Negotiated Payments
____Manual Billing
____Automated Billing

_______**Supervision/Training**
____Clerical Personnel
____Accounting Staff
____Work Delegation
____Departmental Liaison

_______**Computer**
____Data Entry
____CRT
____Wordprocessing
____Spreadsheets
____Database Input
____Formletters/Mail Merge
____Correspondence
____Reports
____Newsletters
____Mailing Lists
____Pricing Lists
____Catalogs

Figure 4: List of jobs available with the skills needed for that position ("Resume Skills And Ability | Officer Manager Resume Skills List Examples Raised Pay $4,800 | Resume skills, Resume skills list, List of skills", 2020)

This above figure shows an example of how the data for the list of positions with the skill set required would look like.



MARY ROWAN

example@example.com | (555) 555-5555 | Miami, FL

SUMMARY STATEMENT

Outgoing Apprentice Hairdresser with willingness to learn and excel in beauty industry. Familiarity with blowouts, coloring and perms. Adept at handling front-office duties, including scheduling appointments and greeting customers.

SUMMARY OF QUALIFICATIONS

- Exceptional knowledge of modern hair trends with outstanding attention to detail and commitment to individualized client care.
- Well-organized record keeper with good schedule management and relationship-building abilities.
- Knowledgeable about keeping clean, well-stocked and properly sanitized stations.

PROFESSIONAL SKILLS

Client Satisfaction

- Increased return customers by 80% through cutting, coloring and styling clients' hair with precision.
- Used scissors, trimmers, clippers, razors and brushes on over 300 customers with 100% customer satisfaction rate.
- Consistently received positive performance reviews from guests.

Hair Consulting

- Assessed customer hair texture, root patterns and facial features to identify optimal styles and offer suggestions.
- Educated customers on products, including salon conditioning treatments and at home care to maximize benefits.
- Offered color swatches to help clients choose desired hair color.

Cleanliness

- Cleaned, maintained and sterilized tools to increase lifetime and cut spending by 30%.
- Sterilized cutting equipment and refilled hair product bottles to assist hair stylists during busy periods.
- Kept work areas, tools and equipment clean and properly sanitized to protect customers from risk of infection.

WORK HISTORY

Hairdressing Apprentice / The Salon Professional Academy - Miami, FL / 05.2017 - Current

EDUCATION

Coral Reef Senior High School - Miami, FL / 2016
High School Diploma

Lilibeth Andrada

Data Analyst

Motivated, teamwork-oriented, and responsible Data Analyst with significant experience in increasing comprehension of reports and presentations by the average professional. Highly educated, possessing a Bachelor's, a Master's, and professional certification in business analytics and statistics. Bilingual in English and Tagalog, with an intermediate understanding of Spanish.

lilibeth@andrada.com 627-513-5070 linkedin.com/in/lilibethandrada github.com/lilibethandrada

TECHNICAL (IT) SKILLS

Data Management	Database Design & Management, Data Quality Assessment, Data Analysis, Pattern & Trend Identification, Visualization of Data Insights
Computer Science	System Administration, Advanced Microsoft Excel Functions, Sharepoint, MySQL, Tableau, Python, Troubleshooting
Research	Data Science Research Methods, Data Mining, Survey Creation, Focus Groups

WORK EXPERIENCE

IT Business Operations Data Analyst
Dell Technologies
04/2014 – Present

- Converted data into actionable insights by predicting and modeling future outcomes.
- Utilized MS SQL, data warehousing programs, Tableau, and other dashboard/visualization toolsets for data intelligence and analysis.
- Successfully adhered to company data governance standards receiving the "Employee of the Year" for the last 2 years in a row.

Contact: John Doe – jd@dell.com

Market Research Analyst
Pearson & Associates
11/2010 – 03/2014

- Identified, analyzed, and executed new and potential products, services, markets, and advertising opportunities.
- Collected and analyzed data on established and prospective customers, competitors, and marketing channels and sources.
- Prepared reports that interpret consumer behavior, market opportunities and conditions, marketing results, trends, and investment levels.
- Utilized Google Analytics and Google Tag Manager and implemented new scripts that increased performance by 25%.

EDUCATION

MBA - Concentration in Business Analytics
University of Texas at Austin
08/2008 – 06/2010

Bachelor's in Statistical Science
Southern Methodist University
08/2005 – 06/2008

PROFESSIONAL CERTIFICATES

Certification of Professional Achievement in Data Sciences (2016)
Online course - Columbia University

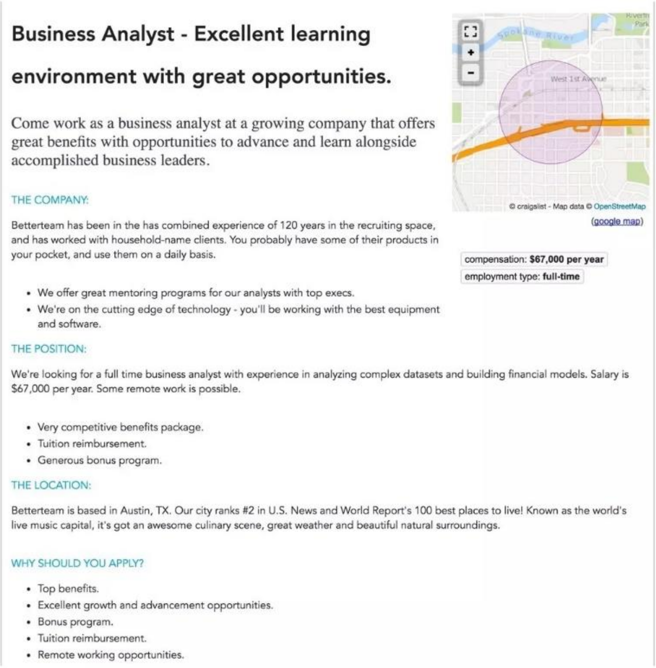
EMC Proven Professional Data Scientist Associate (2014)
Dell EMC Education Services

novoresume.com

Figure 5: Sample resume data

("Check Out The Top Resume Examples For 2020", 2020) (ApS, 2020)

The figure attached above depicts the resume that the user submits through their SEEK portal.



Business Analyst - Excellent learning environment with great opportunities.

Come work as a business analyst at a growing company that offers great benefits with opportunities to advance and learn alongside accomplished business leaders.

THE COMPANY:

Betterteam has been in the has combined experience of 120 years in the recruiting space, and has worked with household-name clients. You probably have some of their products in your pocket, and use them on a daily basis.

- We offer great mentoring programs for our analysts with top execs.
- We're on the cutting edge of technology - you'll be working with the best equipment and software.

compensation: **$67,000 per year**
employment type: **full-time**

THE POSITION:

We're looking for a full time business analyst with experience in analyzing complex datasets and building financial models. Salary is $67,000 per year. Some remote work is possible.

- Very competitive benefits package.
- Tuition reimbursement.
- Generous bonus program.

THE LOCATION:

Betterteam is based in Austin, TX. Our city ranks #2 in U.S. News and World Report's 100 best places to live! Known as the world's live music capital, it's got an awesome culinary scene, great weather and beautiful natural surroundings.

WHY SHOULD YOU APPLY?

- Top benefits.
- Excellent growth and advancement opportunities.
- Bonus program.
- Tuition reimbursement.
- Remote working opportunities.

Figure 6: Sample job posting data
("How To Write a Great Job Posting [Examples and Templates]", 2020).

The above figure shows a sample of the job posting that the hirer posts on SEEK.
Once the data is obtained, it can be processed for further interpretation and analysis according to the location and the profile of the user.
In order to proceed with the project, it is necessary to set up a virtual environment first and have *Jupyter* notebook which has kernels for both Python and R installed in it. Many modules like *numpy*, *pandas*, *scipy*, *NLTK*, *scikit-learn* will be used. Packages like *tm* in R is used for basic text mining and wrangling. *Apache Spark* or *Hadoop* is used for handling large amounts of data present.

DATA ANALYSIS:

In a data science project, decisions are based on data which means with the help of data and its analysis, the project can be a success. The main motive of data analysis is to understand the significance of the data so that it can be helpful in making decisions. For this job recommender system, getting sufficient amount of data is extremely helpful for providing accurate suggestions. The importance of recommendation system is quite significant. The job recommendation system takes into account the user data and the metadata as well. These recommendations based on the users' data directly impact the satisfaction of the job seeker as well as the hirer. One of the simplest way of matching resumes to the job descriptions is by keyword matching. For this, the term frequency – inverse document frequency (TF-IDF) is used to assign more weights to certain keywords that are unique for that particular user and his/her resume.

Artificial Intelligence and machine learning along with natural language processing algorithms can prove to be of more significance for the job recommender system.
As the user submits their resume through their portal, it goes into the data processing block. This is the block where the entire processing of the resume takes place. In order to gain information from the resume after the basic processing and wrangling is performed, syntactic analysis and semantic analysis has to be performed.
Syntactic Analysis: This is the process of analyzing the language with proper grammatical rules. The grammatical rules are used with a group of words and not individual words.
Semantic Analysis: This process tells about the meaning of the sentence from the meaning of its parts.
Python programming language has a package called *Natural Language Toolkit* (*NLTK*) which is used for text processing by tokenization, parsing and semantic reasoning.

After the completion of text processing, a similarity check is carried out by the system. Nearest-neighbour methods are used for calculating the similarity between resumes. The similarities can be calculated as a cosine similarity. A threshold value is kept on the similarity score i.e. if the similarity score is equal to or above that threshold value, then only the resumes can be treated as similar. The matching resumes are then put under the domains available for them to work according to the data provided, which tells about the various jobs and the skill set needed for each position. Matching the skill sets and putting them under the same domain keeping in account the threshold value, can be done using the libraries like *gensim*, *textblob* and *spacy* in Python.
This approach is Collaborative filtering where if user X and user Y are in the same domain and X is recommended job A and Y is recommended job A and B, so with this feature user X is also recommended job B as both the users fall under the same domain.

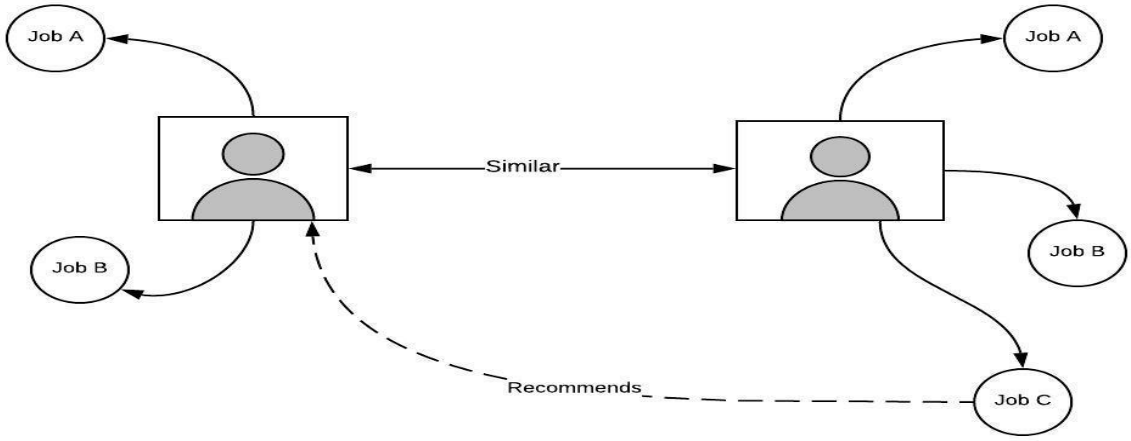


Figure 7: Collaborative filtering

The next step would involve the building of a Multi-class classifier. Multi-class classification involves grouping something (resumes) into one of the classes (jobs). For this a Python module *scikit–learn*, that integrates machine learning algorithms using *numpy*, *pandas*, *matplotlib* can be used. This can be used to provide efficient results to the problems by considering a set of samples and predicting unknown data. For the given dataset, some portion of this is left for testing the model while the rest is used in training the model.

Basic training of the model can be done by using the existing data of the available jobs and the resumes' data by using the machine learning techniques using the python library *keras.*

Many models like the K-Nearest Neighbours, Random Forest, Support Vector Machine are trained according to our data. K-Nearest Neighbours is an algorithm that keeps and stores all the cases and then it classifies the new cases based on the similarity measured. In this case, it stores all the resumes and classifies the domain of the resume based on the similarity value (Subramanian, 2019).

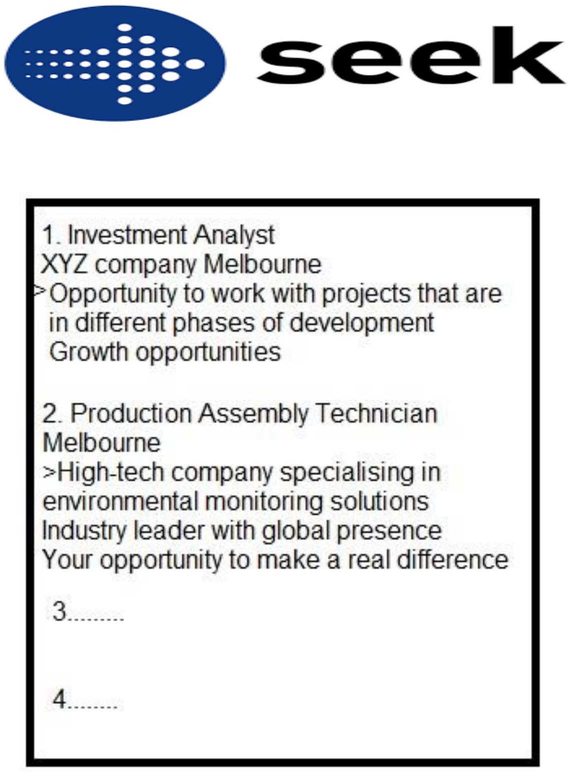
Random Forest classifier is a machine learning algorithm which trains many decision trees and combines the results of all the trees in order to produce a better result.

Support Vector Machine is a supervised machine learning algorithm which is used for classification by doing some transformations of data and then separating the data based on the outputs that have been defined (Sarkar, 2019).

If there is a significant variance in the accuracy over the testing data and the training data, this means that our model is overfitting. Overfitting can be tackled by either reducing the majority instance or increasing the minority instance.

We'll fix a random baseline model which can define the threshold value for the models on which the training is performed. The results of the models are then compared to the baseline model. The model with the maximum accuracy is selected as the best classifier and then the parameters of the model are tuned for getting proper values of the parameter. The simplest way for finding the best values for the parameters can be cross-validation, and understanding which subset of the parameter should be used first as all of the parameters are not needed in a lot of cases to get a more accurate result.

This approach can be used to get the best recommendations for a job seeker based on his/her resume.



1. Investment Analyst
XYZ company Melbourne
>Opportunity to work with projects that are in different phases of development
Growth opportunities

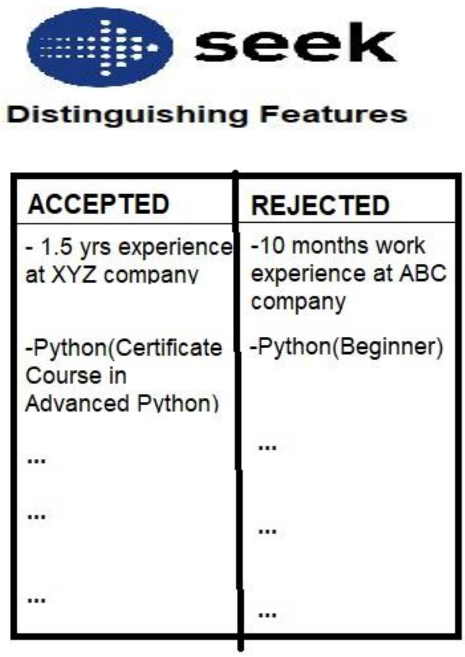
2. Production Assembly Technician
Melbourne
>High-tech company specialising in environmental monitoring solutions
Industry leader with global presence
Your opportunity to make a real difference

3.........

4........

Figure 8: Sample output of recommended jobs for user X

At a later stage, the job seekers who get rejection for the job they applied to get a detailed feedback from SEEK with the help of this model which helps in finding the distinguishing factors between the selected resume and the rejected resume also tells the differences in the skills of the chosen candidate and the rejected candidate.



Distinguishing Features

ACCEPTED	REJECTED
- 1.5 yrs experience at XYZ company	-10 months work experience at ABC company
-Python(Certificate Course in Advanced Python)	-Python(Beginner)
...	...
...	...
...	...

Figure 9: Sample output to show the distinguishing factors

CONCLUSION:

Implementing a data science project is a complex task and requires a lot of people with various skill sets. In this data science project, one of the many applications of artificial intelligence, machine learning, natural language processing and text mining algorithms are used for recommending jobs to job seekers based on their profiles. For this project efforts have been put to take into consideration the skill set of the candidate and also pointing out the skills that the candidate lacks and needs to achieve in order to be eligible for a particular type of position. This job recommender system will provide better and efficient solution to the current hiring system at SEEK thus providing a potential candidate to the company which needs such skillset.

REFERENCES:

- Appan, P. (2016). Algorithms and architecture for job recommendations. Retrieved 10 May 2020, from https://www.oreilly.com/content/algorithms- and-architecture-for-job-recommendations/
- Delgado, M. (2019). How Do People Find Jobs? | Clutch.co. Retrieved 12 May 2020, from https://clutch.co/hr/recruiting/resources/how-people-find-jobs.
- About SEEK. (2019). Retrieved 13 May 2020, Retrieved from https://www.seek.com.au/about/.
- Subramanian, D. (2019). A Simple Introduction to K Nearest Neighbors Algorithm. Retrieved 31 May 2020, from https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e
- Sarkar, P. (2019). Machine Learning: What Are Support Vector Machines(SVMs)?. Retrieved 31 May 2020, from https://www.knowledgehut.com/blog/data-science/support-vector-machines-in-machine-learning
- How To Write a Great Job Posting [Examples and Templates]. (2020). Retrieved 31 May 2020, from https://www.betterteam.com/job-posting-template
- Check Out The Top Resume Examples For 2020. (2020). Retrieved 31 May 2020, from https://www.myperfectresume.com/resume-samples
- ApS, N. (2020). Data Analyst Resume - Guide & Examples for 2020. Retrieved 31 May 2020, from https://novoresume.com/career-blog/data-analyst-resume
- Resume Skills And Ability | Officer Manager Resume Skills List Examples Raised Pay $4,800 | Resume skills, Resume skills list, List of skills. (2020). Retrieved 31 May 2020, from https://in.pinterest.com/pin/111745634476727606/
- skills, C. (2020). Job descriptions: A to Z of careers. Retrieved 31 May 2020, from https://targetjobs.co.uk/careers-advice/job-descriptions
- A Guide to Apache Spark Streaming - Intellipaat Blog. (2017). Retrieved 31 May 2020, from https://intellipaat.com/blog/a-guide-to-apache-spark-streaming-tutorial/
- Gill, J. (2018). Data Preparation,Preprocessing and Wrangling Tools - XenonStack. Retrieved 31 May 2020, from https://www.xenonstack.com/blog/data-preparation/