

Team Name	Zero	
Student ID	Name	E-mail
A0178551X	Choo Ming Hui Raymond	e0267862@u.nus.edu
A0178431A	Huang Qingyi	e0267742@u.nus.edu
A0178415Y	Jiang Zhiyuan	e0267726@u.nus.edu
A0178365R	Wang Jingli	e0267676@u.nus.edu
A0178329R	Wong Yeng Fai, Edric	e0267640@u.nus.edu
A0178371X	Yang Shuting	e0267682@u.nus.edu



NUS
National University
of Singapore



INSTITUTE OF SYSTEMS SCIENCE

KE5205 Text Mining Assignment

**Mining Machine Learning related
job information from mycareersfuture.sg**

Lecturer: Ms. Fan Zhen Zhen

Project Outline

Project Objective:

As we will be working in data-related positions upon graduation, a clear understanding of the current demands of these positions in the job market, and what to expect by working in this field would guide us in making the right decisions to apply for a job, and also prepare us for recruitment process.

Questions to Answer:

In this assignment, text mining techniques were used on data-related jobs posted on mycareersfuture.sg to answer the following questions:

1. What is the market outlook for “machine learning” jobs?
2. Based on an applicant’s interest, what are the job descriptions available to apply for?
3. Based on an applicant’s skillset, what are the minimum required to apply for generic jobs?
4. What are the career paths and opportunities available?

Table of Contents

1.	Background	1
2.	Data Acquisition	1
3.	Textual Data Preparation	3
3.1	From text to words:	3
3.2	POS tagging:	3
3.3	Information Extraction	4
4.	Text Mining Findings	4
4.1	Q1: What is the market outlook for “machine learning” jobs?	4
4.2	Q2: What are the main data-related job titles and the corresponding key requirements? 6	
4.3	Q3: What are the main responsibilities for different jobs?	9
4.4	Q4: What are the career paths and opportunities ahead?	13
5.	Conclusion	19
6.	References	19
	Appendix I	1

1. Background

Artificial Intelligence (A.I.) and Machine Learning (ML) has been developing at an accelerated pace in recent years, gradually disrupting diverse domains such as human interaction, healthcare, Smart Cities and technical industries, to name a few. Facing such a trend, the Singapore government also takes massive efforts in nurturing A.I. capabilities. For instance, the National Research Foundation (NRF) launched the “**A.I. Singapore**” initiative on 30th August this year with an aim to boost Singapore’s artificial intelligence capabilities by enabling 12,000 more people to acquire A.I. knowhow, investing up to \$150 million dollars over five years[1], [2].

Under this macro-environment, the demand for talent in the A.I. and ML fields is also increasing in tandem. As graduates from this field, having a clear understanding in the current demands of the job market is crucial for both career planning, and the preparation of the recruitment process. For this purpose, text mining techniques were applied to extract and analyse the job market from the government’s jobs portal 'mycareersfuture.sg' by searching for the key word '*machine learning*'.

2. Data Acquisition

In this project, the first task was to collect data from the jobs portal 'www.mycareersfuture.sg' where over 20,000 jobs are posted. By searching the '*machine learning*' keyword, a list of 216 job opportunities were obtained.

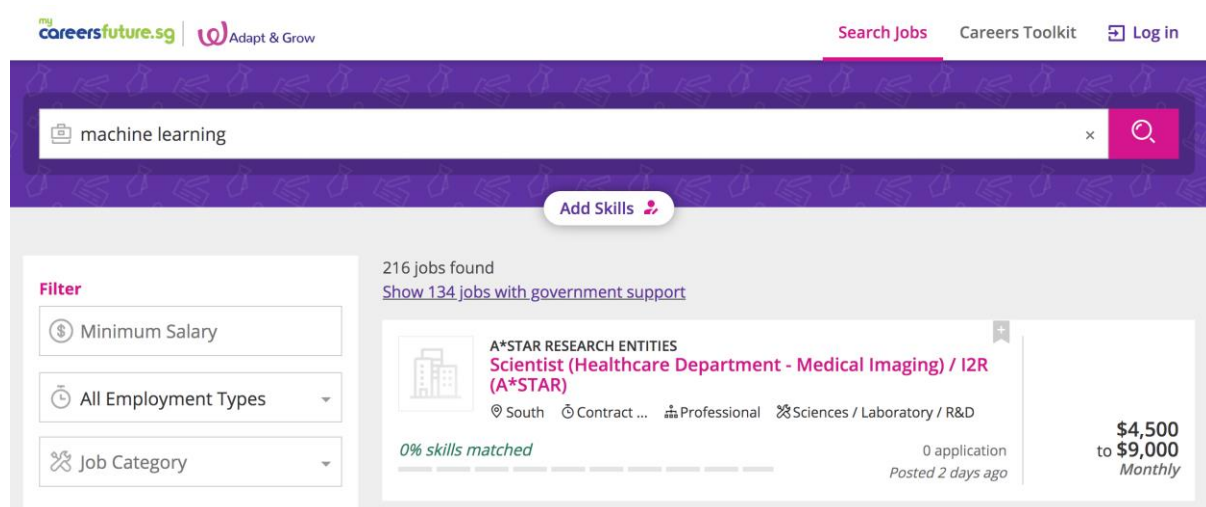


Figure 1: Querying for jobs using “Machine Learning” on mycareersfuture.sg

The results returned were listed in 11 pages, with each page containing 20 job profiles. Intuitively, it is not practical to extract the job details one by one manually. Hence, using the Python packages - *Selenium* and *Beautiful Soup*, an automatic web crawler script was developed to extract the detailed information for each job. Through the job list and details webpages shown in Figure 2, two areas were targeted: (1) Structured data including company, job title, salary, geo location, seniority, etc. (2) Unstructured text data involving job requirement and job description. The captured data was converted into a *dataframe* format and stored into a CSV file. The data dictionary is included in **Appendix I**.

EPS COMPUTER SYSTEMS PTE LTD
Associate Consultant
 Islandwide Contract ... Executive ... Information Technology
 0% skills matched 0 application Posted 5 days ago
 Government support available
\$5,000 to \$7,000 Monthly

EPS COMPUTER SYSTEMS PTE LTD
Associate Consultant
 SPORE BUSINESS FEDERATION CTR, 160 ROBINSON ROAD 068914 Contract, Full Time
 Executive, Senior Executive Information Technology
\$5,000 to \$7,000 Monthly
 0 application Posted 10 Oct 2018 Closing on 09 Nov 2018

Roles & Responsibilities

- Identify valuable data sources and automate collection processes
- Undertake preprocessing of structured and unstructured data
- Analyze large amounts of information to discover trends and patterns
- Build predictive models and machine-learning algorithms
- Combine models through ensemble modeling
- Present information using data visualization techniques
- Propose solutions and strategies to business challenges
- Collaborate with engineering and product development teams

Requirements

- Proven experience as a Data Scientist or Data Analyst
- Experience in data mining
- Understanding of machine-learning and operations research
- Knowledge of R, SQL and Python; familiarity with Scala, Java or C++ is an asset
- Experience using business intelligence tools (e.g. Tableau) and data frameworks (e.g. Hadoop)
- Analytical mind and business acumen
- Strong math skills (e.g. statistics, algebra)
- Problem-solving aptitude
- Excellent communication and presentation skills
- BSc/BA in Computer Science, Engineering or relevant field; graduate degree in Data Science or other quantitative field is preferred

Figure 2: Job details introduction. Red grid indicates content targeted by web crawler

3. Textual Data Preparation

Prior to mining data, pre-processing is essential to extracting meaningful information. The following sections summarises the entire pre-processing process.

3.1 From text to words:

1. Tokenisation: split text into words
2. Case lowering: standardise words into lower case
3. Punctuation removal (not for Collocation purpose)
4. Stop words removal (not for Collocation purpose)
5. Lemmatisation (not for Collocation purpose)

3.2 POS tagging:

In our case, two raw text data columns '*job_requirement*' and '*job_description*' are unstructured containing many useless words. To refine and extract more informative words out of these columns, POS tagging was applied by which hints from examining the grammatical structure and parts of speech can be obtained. The Stanford CoreNLP toolkit was used to visualise the POS tags of each word in a sentence shown in Figure 3.

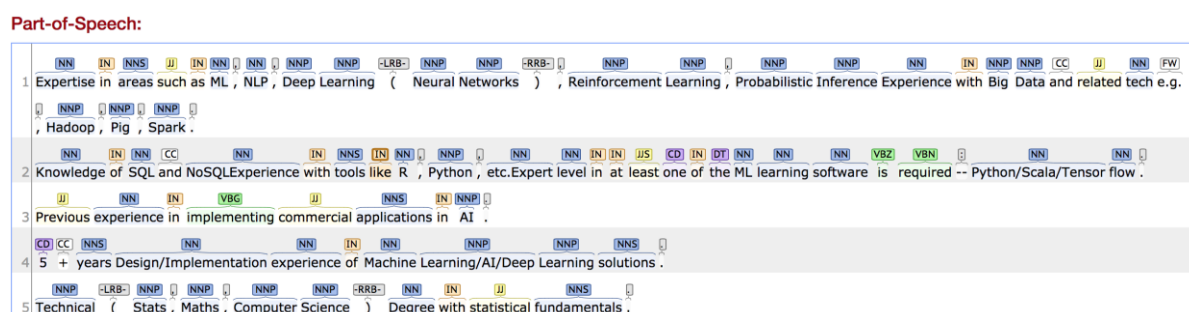


Figure 3: POS tagging using Stanford CoreNLP toolkit

It was determined that Nouns (tagged by NN, NNP, NNS, NNPS) and Verbs (tagged by VBZ, VBN, VBG) would carry more information in sentences. As such, words tagged by NN, NNP and VBZ were extracted and stored in new columns for further mining analysis to reduce noise for information extraction.

3.3 Information Extraction

Narrowing the focus down on two specific requirements, (i) working experience and (ii) education level, the following rules were defined and applied to extract key information using Regular Expression:

- Years of Working Experience:

$[\backslash d] * [\backslash +]? \text{year}[s]? [^.] * [\backslash . \backslash ; ;]$

- Education Requirements:

Bachelor|Master|PhD|Doctor|Ph. D|Diploma|Masters|BS|MS|BA|PHD

As such, new features containing extracted information were generated and shown below in Table 1:

Table 1: Extracted features

S/N	Variable	Description	Value
1	title	Job title	junior, senior, manager
2	job_requirement_nn	Extraction of all nouns in job requirement	
3	job_description_nn	Extraction of all nouns in job description	
4	job_requirement_nnp	Extraction of all proper nouns in job requirement	
5	job_description_nnp	Extraction of all proper nouns in job description	
6	job_requirement_vb	Extraction of all verbs in job requirement	
7	job_description_vb	Extraction of all verbs in job description	
8	job_experience	Working experience requirement	numerical
9	job_degree	Education level requirement	PhD, master, bachelor, none

4. Text Mining Findings

4.1 Q1: What is the market outlook for “machine learning” jobs?

Overview. All 216 records derived from section 2 and 3 were examined in terms of the company recruiting, the location of these companies, and the respective industries they represent. With these information, the recruitment demand for all 21 industries were ranked,

and the location of these jobs and the scale of recruitment from the recruiting company are shown in Figure 4.

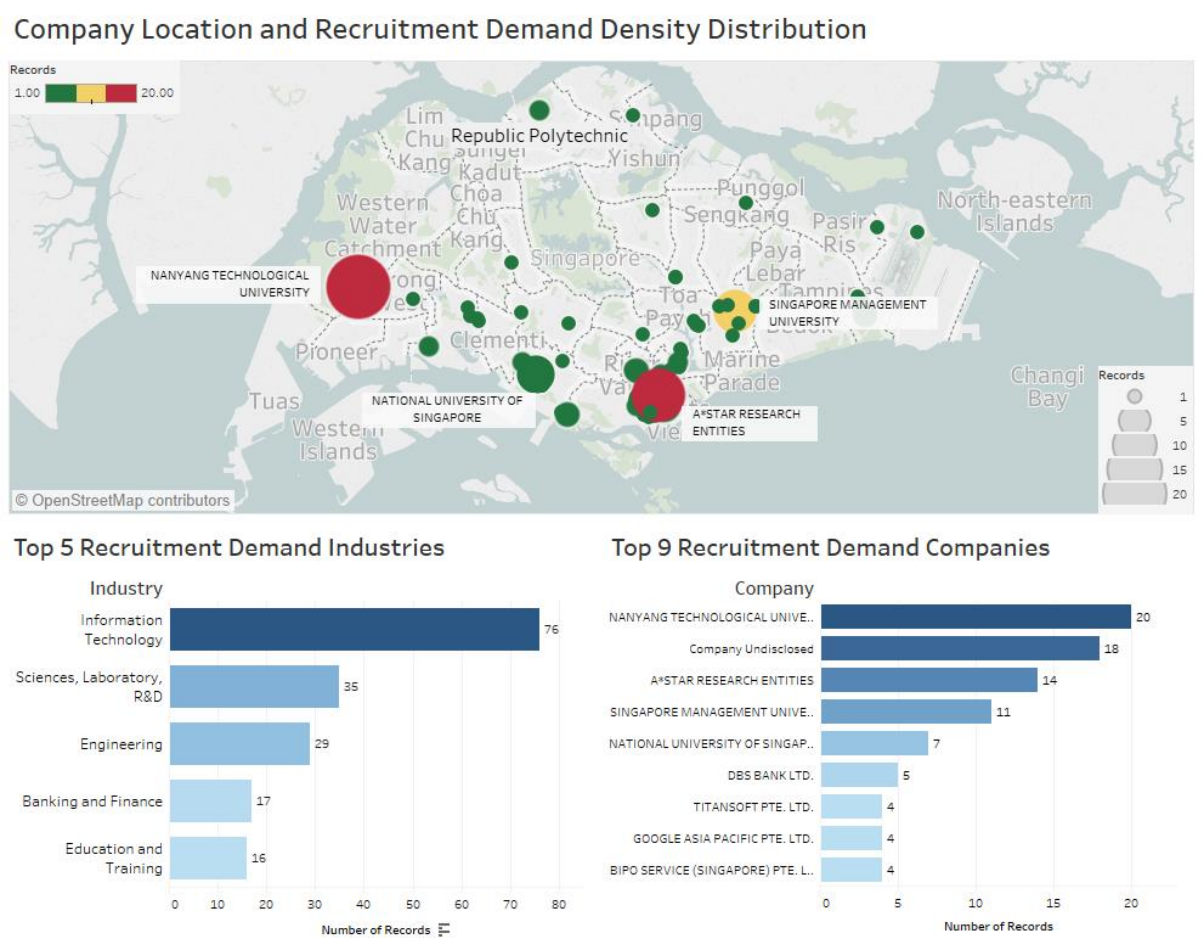


Figure 4: Overview of market demand for “Machine Learning”

Insights. From the company location and recruitment demand density distribution, the nodes represent the distinct recruiting companies, whereas the size and colour of the nodes correspond to the number of jobs provided by each company. It is noted that most companies with a high demand for “machine learning” roles are located in the south-west and central part of Singapore. From the Top 5 Recruitment Demand Industries, it is also observed that there are more job opportunities in the “*Information Technology*”, “*Science, Laboratory, R&D*”, “*Engineering*”, “*Banking and Finance*” and “*Education and Training*” industries. The “*Information Technology*”, “*Science, Laboratory, R&D*” industries collectively account for nearly half of the entire demand, which signifies that these are the hottest industries currently for this query. In addition, from the top 9 recruitment demand companies, it is noted that the recruitment demand of the top three public universities in Singapore and A*STAR for “machine learning”

talents are the highest, reflecting the commitment of the government of Singapore in nurturing and providing an avenue for talent and career development, which is in line with the discussion on “*AI Singapore*” in section 1.

4.2 Q2: What are the main data-related job titles and the corresponding key requirements?

Overview. The *‘job_requirement’* column contains key job requirements for each job, such as basic skills, working experience and education levels. To answer question 2, the K-means clustering technique was applied to group jobs with similar requirements into the same cluster, with the interpretation that each cluster corresponds to a job title.

Mining Approach. Prior to clustering, text pre-processing was carried out to transform raw text to single words. In this stage, the top 200 words were sorted by frequency, and among these terms, meaningless words such as *‘a*star’*, *‘singapore’*, and *‘data’* were added into the stop word list where words will be excluded in further analysis. Then, a term frequency matrix was generated based on TF-IDF indexing. Since the size of the matrix was only 216×1070, dimension reduction was ruled out. This derived matrix was thus fitted directly into a K-means clustering model. By compared the clustering results, it is noted that jobs were well classified when the pre-defined number of clusters was chosen to be 5. As K-means is susceptible to reaching local optima, it took several runs for the algorithm to converge a global optimum. To further improve the clustering performance, column *‘job_requirement_nn’* where only nouns are remained was used for analysis instead. In addition, in order to gain a better understanding of different positions, following mining methods were conducted to extract requirements from different aspects:

1. To determine cluster names (job title), closest words to each cluster center and high frequency terms and phrases extracted by unigram and bigram were considered together.
2. To identify the working experience requirement, aggregation was applied on column *‘job_experience’*.
3. To identify education level requirement, aggregation was applied on column *‘job_degree’*.
4. To summarize key required skills, unigram and bigram frequency count was conducted on column *‘job_skills’*

5. To extract software skills requirement, unigram and bigram frequency count was conducted on column '*job_requirement_nnp*' where proper nouns were retained.

Insights. The clustering results are visualised in Figure 5 below, with 5 main data-related job titles proposed:

1. Business Manager
2. Business Analyst
3. Researcher & Assistant
4. Data Engineer
5. Developer

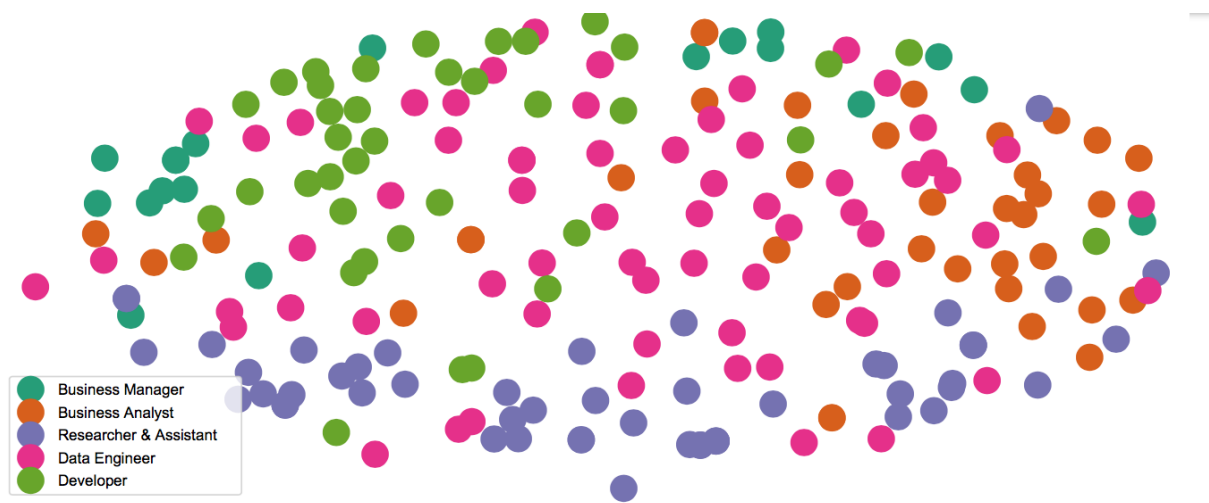


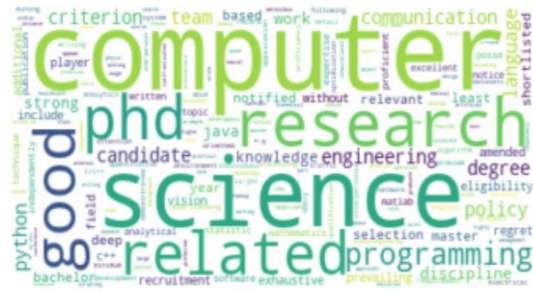
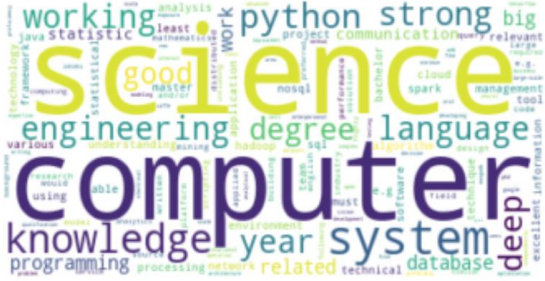



Figure 5: K-means Clustering for “machine learning” jobs, K = 5

Table 2 illustrates the key requirements for each job title, such as working experience, education level, key skills and software skills.

Table 2: Summary of Job Requirements

Job Title	Key Requirements
<p><u>Business Manager</u></p> <p>19 positions available; 9% of total jobs available</p>  <p>A word cloud for 'Business Manager' with 'business' as the largest word. Other prominent words include 'team', 'technology', 'strong', 'management', 'working', 'role', 'client', 'service', 'communication', 'solution', 'leader', 'transformation', 'value', 'people', 'knowledge', 'technical', 'process', 'dynamic', 'complex', 'develop', 'company', 'support', 'opportunity', 'apply', 'science', 'industry', 'join', 'make', 'work', 'model', 'system', 'passionate', 'excellence', 'degree', 'communication skills, management ability, strategic planning', 'software skills', 'none'.</p>	<ol style="list-style-type: none"> 1. Working Experience: Working experience is crucial to this position Preferred >5 years working experience 2. Degree: No compulsory requirement 3. Key Skills communication skills, management ability, strategic planning 4. Software Skills None
<p><u>Business Analyst</u></p> <p>38 positions available; 18% of total jobs available;</p>  <p>A word cloud for 'Business Analyst' with 'analysis' as the largest word. Other prominent words include 'degree', 'business', 'python', 'analytics', 'statistical', 'science', 'communication', 'statistic', 'knowledge', 'tool', 'working', 'team', 'programming', 'project', 'management', 'security', 'excellent', 'demonstrated', 'qualifications', 'strong', 'engineering', 'technology', 'computer', 'related', 'software', 'quantitative', 'understanding', 'industry', 'mission', 'mathematical', 'relevant', 'model', 'research', 'discipline', 'big', 'data', 'master', 'equivalent', 'development', 'track', 'preferred', 'project', 'management', 'communication', 'statistic', 'statistical', 'knowledge', 'player', 'team', 'player, statistical knowledge', 'microsoft office, python, spark, sql'.</p>	<ol style="list-style-type: none"> 1. Working Experience: Preferred 3-5years working experience 2. Degree: Preferred Master's Degree or above 3. Key Skills communication skills, project management, team player, statistical knowledge 4. Software Skills Microsoft Office, Python, Spark, SQL
<p><u>Researcher & Assistant</u></p> <p>55 positions available; 25% of total jobs available;</p>  <p>A word cloud for 'Researcher & Assistant' with 'computer' as the largest word. Other prominent words include 'research', 'science', 'related', 'programming', 'communication', 'team', 'based', 'work', 'criterion', 'player', 'include', 'strong', 'phd', 'candidate', 'knowledge', 'engineering', 'degree', 'eligibility', 'policy', 'regret', 'discipline', 'python', 'java', 'c++', 'matlab', 'deep', 'field', 'vision', 'analysis', 'recruitment', 'software', 'exhaustive', 'prevailing', 'discipline', 'selection', 'master', 'with', 'regret', 'discipline', 'python, java, c++, matlab', 'artificial intelligence, programming ability'.</p>	<ol style="list-style-type: none"> 1. Working Experience: No compulsory requirement 2. Degree: High level degree is crucial to this position Preferred PHD, or Master Degree in computer science 3. Key Skills artificial intelligence, programming ability 4. Software Skills Python, Java, C++, MATLAB

<p style="text-align: center;"><u>Data Engineer</u></p> <p style="text-align: center;">66 positions available; 31% of total jobs available;</p> 	<ol style="list-style-type: none"> 1. Working Experience: No compulsory requirement, preferred 3-4 years working experience 2. Degree: Preferred Bachelor's Degree or above in Computer Science 3. Key Skills communication skills, artificial intelligence, large scale database, deep learning 4. Software Skills Python, Java, Spark, Hadoop, NoSQL, tensor flow
<p style="text-align: center;"><u>Developer</u></p> <p style="text-align: center;">38 positions available; 18% of total jobs available;</p> 	<ol style="list-style-type: none"> 1. Working Experience: Preferred > 5 years working experience 2. Degree: Preferred Bachelor's Degree or above in computer science 3. Key Skills Software development, communication skills, team players, artificial intelligence, programming 4. Software Skills Java, Python

4.3 Q3: What are the main responsibilities for different jobs?

Overview: The answer of this question focuses on the ‘*job_description*’ column, which records the main responsibility or role in a long raw text. Through this column, the company business, scope of the job or the A.I. project topic could be mined from the raw text data using word-term statistics and the Latent Dirichlet Allocation (LDA) method. In addition, from the results from the LDA method, the various topics distribution and ratio can be visualized by visualization tools.

Mining Approach: It is observed that the raw text data of the ‘*job_description*’ column contains many meaningless POS-tag words. In order to extract more informative words and phrases, 3 kinds of n-grams corpus were created: unigram, bigram, mixgram (combine unigram and

bigram). As such, the word-term frequency of mixgram and bigram were calculated showing in Figure 6.

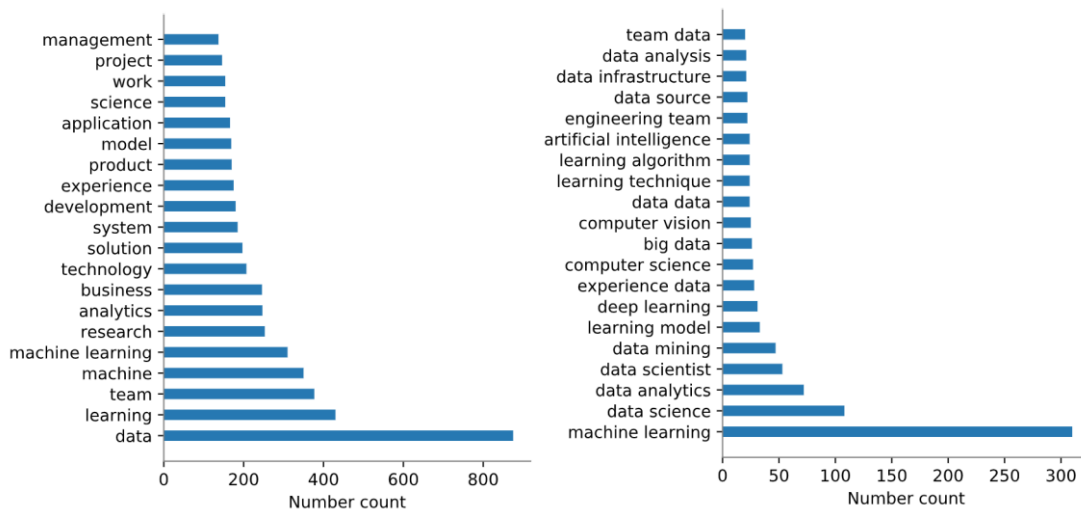


Figure 6: Word-term frequency ranking list for mixgram and bigram

In the ranking lists, it was observed that the word-frequency has a long-tail distribution. From high frequency words like “data”, “team”, “learning”, “research” and the high frequency phrases like “machine learning”, “data science”, “data analytics”, the job responsibilities revolve around data related works with machine learning methods. From the results of the middle part of the ranking list, it is noted that there are a few words which may appear to have summarised the different topics such as “business”, “management” or “computer vision”. However, the occurrences of these words are not numerically significant to base our conclusion on these terms.

To extract the key words for different topics, the LDA model was introduced to allow sets of observations to be explained by unobserved groups that explain why some parts of documents are similar. By doing this, every document of job description would belong to corresponding groups. In our case, the mixgram corpus was firstly used to build a dictionary. Subsequently, the TF-IDF model was applied to convert the words to vectors. Finally, the LDA module *gensim* was invoked to generate the LDA results in Table 3. There were several configurations that has to be set in advance such as the number of topics (set as 8), number of key words (set as 8) and upper-bound of word frequency (set as 0.5). The visualization of LDA was generated using the *pyLDavis* package as shown in 7.

Table 3: Group topics with LDA model

Group	Ratio	LDA Expression	Topic
1	26.0%	0.006*"experience" + 0.005*"product" + 0.005*"code" + 0.005*"model" + 0.004*"system" + 0.004*"client" + 0.004*"lab" + 0.004*"service"	Develop product or service for clients
2	16.8%	0.005*"research" + 0.004*"ai" + 0.004*"processing" + 0.004*"experience" + 0.004*"security" + 0.004*"language" + 0.004*"network" + 0.003*"application"	Research NLP or networking security
3	14.0%	0.005*"analytics application" + 0.005*"analytics" + 0.005*"solution" + 0.005*"analysis" + 0.004*"management" + 0.004*"project" + 0.004*"process" + 0.004*"capability"	Solution through analytics
4	12.2%	0.006*"client" + 0.005*"computer vision" + 0.005*"vision" + 0.004*"service" + 0.004*"marketing" + 0.004*"analytics" + 0.004*"product" + 0.004*"business"	Computer vision with marketing
5	11.5%	0.005*"process" + 0.005*"research" + 0.005*"developer" + 0.004*"ai" + 0.004*"development" + 0.004*"software" + 0.004*"algorithm" + 0.004*"fusion"	Data fusion development and research
6	8.3%	0.005*"region" + 0.005*"research" + 0.005*"data science" + 0.005*"behaviour" + 0.004*"customer" + 0.004*"user" + 0.004*"software" + 0.004*"vehicle"	Customer behavior, vehicle related reseach
7	5.9%	0.008*"traffic" + 0.006*"research" + 0.006*"nanyang" + 0.006*"school" + 0.005*"nanyang technological" + 0.005*"technological" + 0.005*"technological university" + 0.004*"company"	Traffic related research in NTU
8	5.2%	0.006*"google" + 0.004*"business" + 0.004*"data analysis" + 0.004*"course" + 0.004*"customer" + 0.004*"research" + 0.004*"information" + 0.003*"engineering"	Data analytics and research in Google

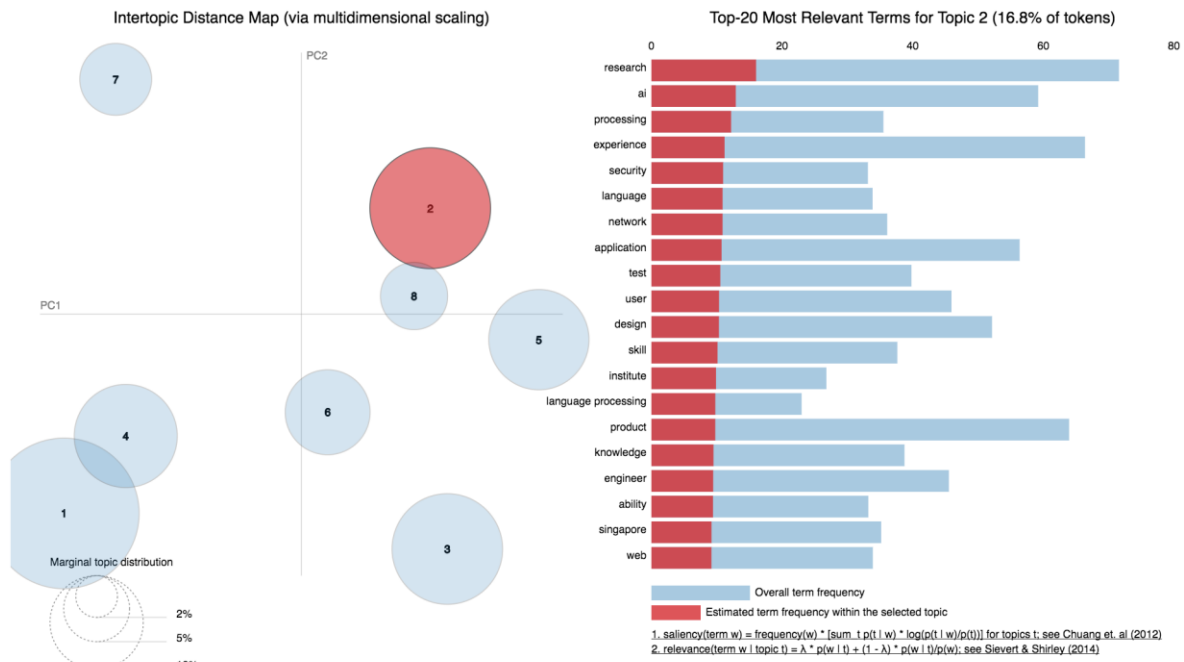


Figure 7: Visualisation result for LDA using pyLDAvis package

Insights: From the visualisation results, it is observed that the topics according to job descriptions are classified under 8 independent groups. These 8 groups could be further summarized into 3 major categories according to bottom-to-top rule, which were *development-orientation*, *analytics-orientation* and *research-orientation*.

For jobs that were development oriented (Groups 1 and 5, consisting of 37.5% of the total jobs available), words such as “*the development*”, “*software*”, “*application*”, “*service*”, etc were as the key points. Based on these job descriptions, the expectation is that these candidates would be responsible for software development, A.I. related application deployment, data fusion and algorithm deployment.

For jobs that were analytics oriented (Groups 3, 4 and 8, consisting of 31.4% of the total jobs available), words such as “*solution*”, “*data analytics*”, “*customer behaviour*”, “*market*”, “*business*” etc were considered as the key points. In this category, candidates are expected to take on roles related to data analytics in the domains of customer behaviours, marketing or business, delivery solutions to clients and management experience.

For jobs that were research oriented (Groups 2, 6 and 7, consisting of 42.5% of the total jobs available), words such as “*computer vision*”, “*natural language processing*”, “*traffic*”, “*customer behaviour*”, “*network security*” etc were treated as the prevailing job content. In

this category, the employers mostly belong to research institutes (such as NTU, A*STAR) and offer the research topics including computer vision, nature language processing, traffic research and network security.

4.4 Q4: What are the career paths and opportunities ahead?

Overview: As a fresh graduate, one generally has two career options: (i) a professional role or (ii) an entry level management role. To understand the feasibility and prospects available for these paths, available roles must first be segregated into distinct groups corresponding to the different stages of a career, shown in Figure 8. Based on these groups, text mining techniques such as information extraction and term frequency were used to derive insights related to industry demand, salary ranges and skillsets required.

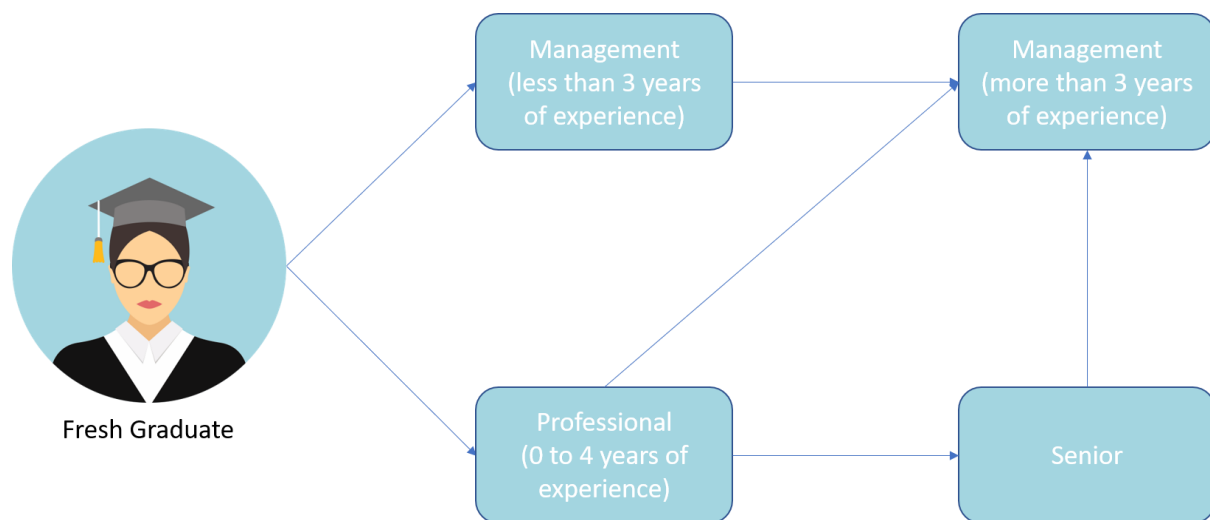


Figure 8: Career Path for a Fresh Graduate

Mining Approach: Initial data exploration reveals that the “*seniority*” field included in mycareersfuture.sg was not consistent. For example, the “Professional” label included managers and directors when these should be classified under “Management”, “Junior Management”, “Senior Management” etc. To classify the dataset into the 4 distinct groups shown in Figure 8, the “*job_experience*” and “*title*” columns created in section 3.3 was used.

Firstly, job titles that contains the management level like terms such as “manager, lead, VP etc” were identified. Based on the “*job_experience*” variable, roles that require less than 3 years of

experience were classified as “Management (less than 3 years of experience)”. Similarly, roles that required more than 3 years of experience were classified as “Management (more than 3 years of experience)”. For records where “*job_experience*” were not indicated, these were excluded from analysis as there is no clear-cut approach to classify these into the 2 groups defined.

With the management roles identified, the next step was to identify the “Senior” roles. These are defined as (i) Job titles with “Senior” indicated; (ii) Any remaining unclassified roles that requires 5 or more years of experience.

The remaining unclassified roles, which are understood to be non-management roles requiring 0 to 4 years of experience are considered suited for fresh graduates to apply and are classified as “Professional”. However, there is a possibility that roles requiring 5 or more years of experience are categorized here due to the “*job_experience*” field left blank. As such, outliers (defined as the mid-point of the salary range greater than \$15,000) for each record were identified and removed.

With these jobs segregated into these groups, the distribution of jobs was evaluated, and the job distribution by industry was done for the entry level management and professional role. The salary profile for each group was also studied using a boxplot, and the skills required were investigated using a word cloud, unigram, bigram and trigram term frequency.

Insights:

Table 4: Distribution of Jobs Across Categories

Management Roles (0 - 3 years of experience)	Management Roles (more than 3 years of experience)
6 Positions (3.1%)	14 Positions (7.3%)
Professional Roles	Senior Roles
131 Positions (67.9%)	42 Positions (21.8%)

1. Distribution of Jobs Across Categories

With reference to Table 4 and Figure 8, it is observed that overall, management positions are limited when compared to the entire spectrum of jobs available. In addition, there are very limited jobs available for jobseekers aiming to be a manager upon graduation. This is consistent with the general knowledge in the distribution of managers and employees in companies, where companies are typically bottom heavy with workers, with a smaller number of managers at the top. It is hence recommended for a fresh graduate to pursue his / her career as a Professional first before aiming towards management, either directly or through Senior roles.

2. Industries in Demand

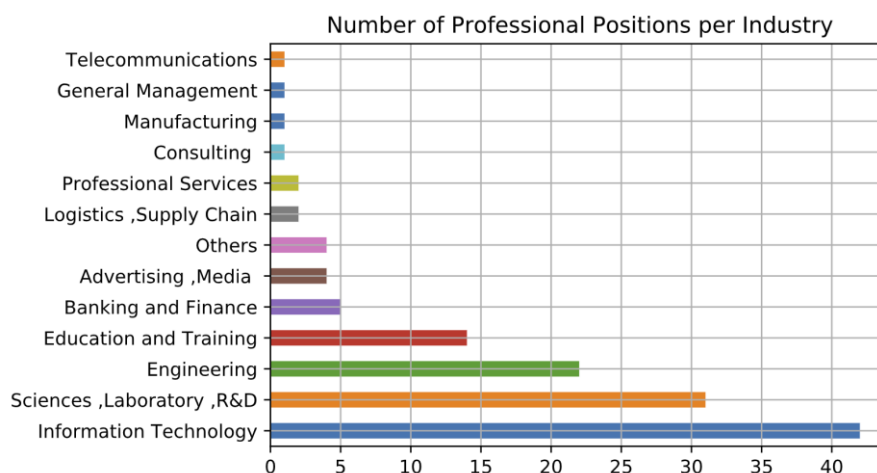


Figure 9: Professional Positions by Industry

For fresh graduates intending to start their career as a Professional, the top 4 industries that are sourcing for talent are from the Information Technology (IT), Sciences / Laboratory / R&D, Engineering, and Education and Training industries. These jobs account for approximately 84% of available jobs for this segment.

3. Salary for Professional Roles

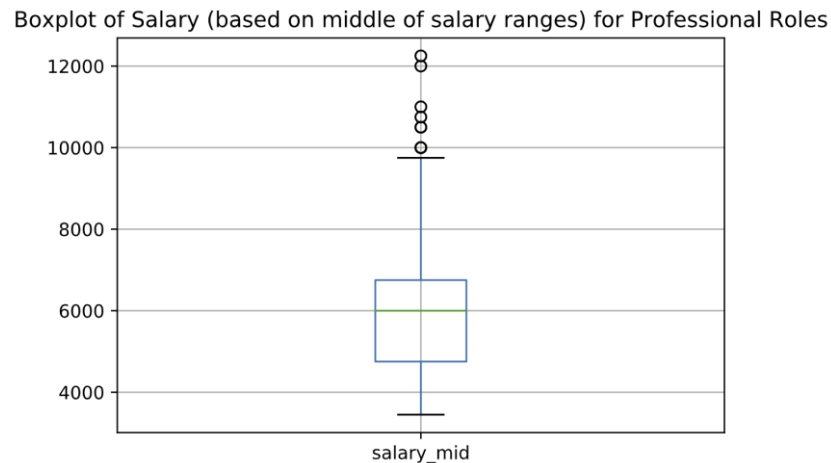


Figure 10: Boxplot of Salary for Professional Roles

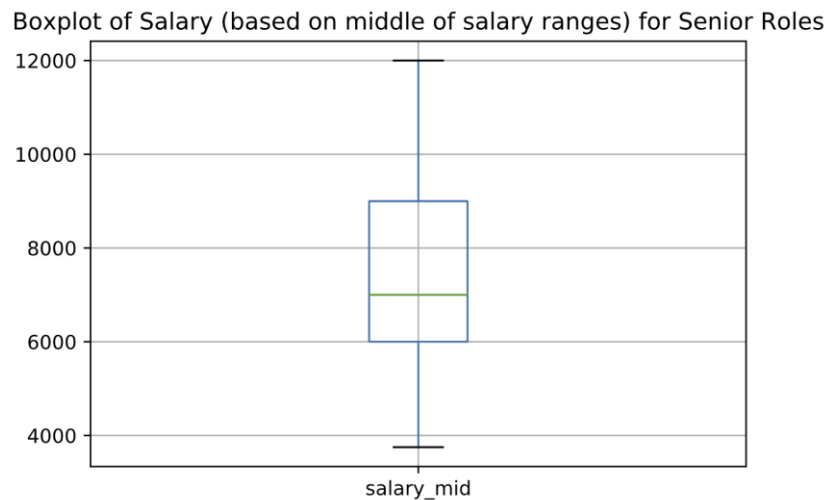



Figure 11: Boxplot of Salary for Senior Roles



A boxplot showing the distribution of 'salary_mid' for the 'none' category. The y-axis ranges from 0 to 70,000 with major grid lines every 10,000. The box is blue, with a median line at approximately 12,000. The interquartile range (IQR) is from about 10,000 to 13,000. Whiskers extend from approximately 8,000 to 16,000. A single outlier is plotted as an open circle at 70,000.

Fresh graduates intending to start their career as a Professional can currently expect their salary to be between \$5,000 and \$6,500, with a median salary of \$6,000. With 4 years of experience under his belt, he can consider pursuing either a Senior role or a Management role. For a Senior Role, he can expect to be paid between \$6,000 and \$9,000, with a median salary of \$7,000 based on today's job postings. Similarly, he can expect to be paid between \$10,000 and \$13,000, with a median salary of around \$12,000 based on today's job postings.

[illegible]

Page 17 of 19

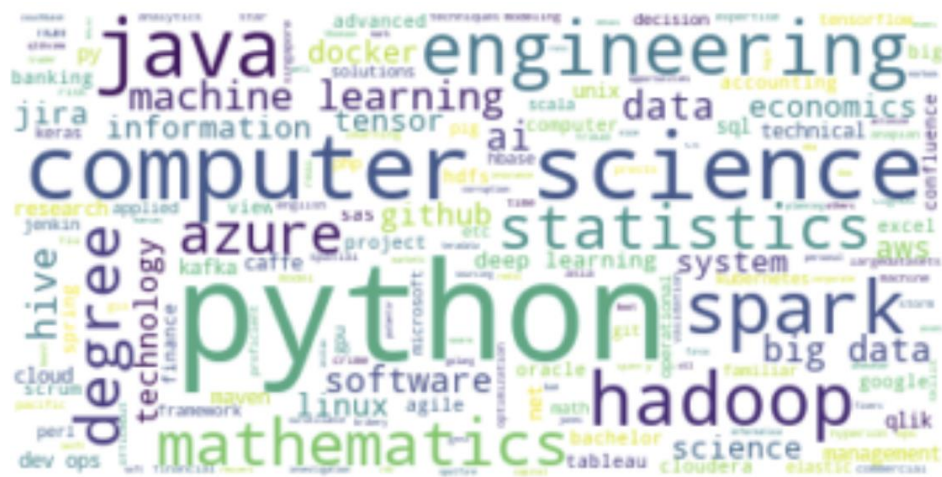


Figure 14: Skills Required for Senior Roles



Figure 15: Skills Required for Management Roles (more than 3 years of experience)

Fresh graduates intending to start their career as a Professional would find it easier to apply for jobs if they are from the Computer Science Engineering background. Skills such as Machine Learning, Python, Java, Spark, Hadoop, Deep Learning etc are highly in demand. If a Professional is looking to progress to the Senior level, it is seen that Computer Science Engineering background, Python, Java, Hadoop and Spark continues to be relevant skillsets required.

Even if one wishes to move onto the Management level, one is still expected to have knowledge of technical skills such as Python and SQL. The only difference is that management roles appear to require travel to, or management of countries such as China, Vietnam, Thailand etc. Management roles also require management skills like

teamwork and interactions with the client in addition to the above. It is thus important for a fresh graduate to grasp their technical skills such as Python well in their early years, as these skills are critical for success in the later stages of their career.

5. Conclusion

With the increasing popularity in the A.I. and ML field, it is inevitable that potential entrants to this field would have questions revolving around the employability and the opportunities available upon graduation and beyond. Through mining the 'mycareersfuture.sg' job portal, 216 jobs related to “machine learning” was captured and analysed. In terms of the market outlook, it was noted that currently, the “*Information Technology*”, “*Science, Laboratory, R&D*”, “*Engineering*” industries have the highest demand for such talent, with the top three public universities in Singapore and A*STAR sourcing for the most personnel. Using K-means clustering, job requirements were studied and the jobs available in the market now are broadly classified as “*Business Manager*”, “*Business Analyst*”, “*Researcher & Assistant*”, “*Data Engineer*” and “*Developer*”. Using LDA modelling, it was also observed that job responsibilities are currently *development*, *analytics* and *research* oriented. In terms of the career path, it is recommended that fresh graduates avoid entry level management roles in the early part of their career due to limited opportunities, and instead focus on technical skills such as Python and Java due to the important of these skills throughout their career. It is hoped that with these findings, individuals seeking to start career in A.I. or ML are better prepared not only for the recruitment phase, but beyond.

6. References

- [1] National Research Foundation (NRF), “AI Singapore.” [Online]. Available: <https://www.nrf.gov.sg/programmes/artificial-intelligence-r-d-programme>. [Accessed: 21-Oct-2018].
- [2] Infocomm Media Development Authority (IMDA), “AI Singapore launches two new initiatives in partnership with IMDA to strengthen local proficiency in Artificial Intelligence (AI).” [Online]. Available: <https://www.imda.gov.sg/about/newsroom/media-releases/2018/ai-singapore-launches-two-new-initiatives-in-partnership-with-imda>. [Accessed: 21-Oct-2018].

Data Dictionary for Original Dataset and Newly Created Features**Table I-1: Original Variables from Dataset**

Column name	Data type	Value
category	string, 28 categories	Engineering, Education and training, Banking and finance and other 25 categories.
company	string	e.g. A*STAR RESEARCH ENTITIES, CITIBANK N.A.
employment_type	string, 3 categories	Full Time, Contract, Permanent
job_title	string	e.g. Data Engineer, Senior Data Analyst, Ad Sales Manager
location	string, 7 categories	Central, West, Islandwide, East, South, North, NaN
latitude	float	e.g. 1.300213
longitude	float	e.g. 103.837286
salary_max	integer	e.g. 3000
salary_min	integer	e.g. 6000
job_requirement	long string	e.g. Skills Required 2+ years of experience. Education in Computer Science, Mathematics, Data Mining, Analytics, Data Science or other quantitative disciplines. Demonstrated aptitude and understanding of modern programming languages with a willingness to continually learn new languages and data structures. Demonstrated interest in computer security issues and current the current cybersecurity threat landscape. Strong analytical, and data analysis skills. Skills desired. Existing knowledge of data analysis tools including Splunk, Elastic. Search, Hadoop. Willingness to learn about the technology and cyber threat environment. General understanding of Cyber security practices. International

		experience or experience working for a global organization.
job_description	long string	e.g. Rolls-Royce@NTU Corporate Lab is currently looking for a candidate to join them as a Research Fellow. The successful applicant will be responsible for: Develop machine learning, AI and optimization algorithms to address real-world problems given by Rolls-Royce. Engage Rolls-Royce stakeholders to capture requirements, formulate research problems, and present findings. Test and evaluate the algorithms, identify their weaknesses and improve them. Guide junior researchers in the team. Prepare reports and scientific papers.

Table I-2: Created Features

S/N	Variable	Description	Value
1	title	Job title	junior, senior, manager
2	job_requirement_nn	Extraction of all nouns in job requirement	
3	job_description_nn	Extraction of all nouns in job description	
4	job_requirement_nnp	Extraction of all proper nouns in job requirement	
5	job_description_nnp	Extraction of all proper nouns in job description	
6	job_requirement_vb	Extraction of all verbs in job requirement	
7	job_description_vb	Extraction of all verbs in job description	
8	latitude	Company location latitude	
9	longitude	Company location longitude	
10	job_experience	Working experience requirement	numerical
11	job_degree	Education level requirement	PhD, master, bachelor, none