

WIU Project Proposal

**19 December 2022**

**TEAM MEMBERS**

|  |  |  |
| --- | --- | --- |
| **Name of Team Members** | **Admin Number** | **Roles & Responsibilities** |
| Zhang Xiang | 210896X | Group Leader |
| Liao Yuan Xia Joanne | 214350B | Group Member |
| Ulysse Yeo Kai Ee | 213903P | Group Member |
| Ng Jing Yi, Clara | 201379H | Group Member |

**Module Group: ITBW51-01**

**Group Name: ANYHOW**

**Project Title: Suitable Job Classifier**

**Tutor: Ms. Jane Zhang**

**Table of Contents**

[**ABSTRACT**](#_heading=h.1cx0suhbd1fa)3

[**INTRODUCTION / BACKGROUND INFORMATION**](#_heading=h.v1myn3nd0kka)3

[**BUSINESS SCENARIO**](#_heading=h.t39ndra80t9e)4

[**BUSINESS OBJECTIVES IDENTIFIED**](#_heading=h.x858dh1uqbg6)4

[Business Objective](#_heading=h.gvv6mivrwmpn) 4

[Measurable Success Criteria](#_heading=h.5vwmkpnowy22) 4

[**MAIN AND SUB TASKS IDENTIFIED**](#_heading=h.pnvvgokpi2ps)4

[**DELEGATION OF TASKS / RESPONSIBILITIES**](#_heading=h.5okg1g24jkuf)5

[Determine Business Goals:](#_heading=h.5gwnu0qpox0y) 5

[Data Collection:](#_heading=h.x66pdvisi2ty) 5

[Data Understanding:](#_heading=h.90tysnts954q) 5

[Data Preparation:](#_heading=h.jyysma4n569k) 5

[Modeling:](#_heading=h.cygpod9xegql) 5

[Evaluation:](#_heading=h.40eszq6ml9zx) 5

[Presentation:](#_heading=h.oj6n3vg4oqri) 5

[**REFERENCE**](#_heading=h.45kkistd3498)7

[Description](#_heading=h.8nx9voegnm7f) 7

[Scraping Website](#_heading=h.fa9zypd75pwn) 7

[Python Library](#_heading=h.jzcnemim9yy3) 7

[Application](#_heading=h.k76217fsm5t1) 7

[**Timeline**](#_heading=h.zqeptbdsxlo)8

[Gantt Chart](#_heading=h.nv7k0kpetacv) 8

[Link](#_heading=h.83c242ajm25i) 8

### ABSTRACT

Looking for a suitable job and in search of eligible candidates are paramount to job seeking and human resource agencies. With extensive information about job descriptions, both employees and employers require assistance to automatically identify job roles based on job description texts. In this proposal, we plan to try using a multi-label classification method to predict job roles such as Data Analyst, Data Scientist, Software Engineer, and Business Analyst by analyzing job description texts and utilizing different classification models. However, this approach will only be attempted if we successfully complete a single label classification approach and if there is enough time available.

This project is to help job candidates in finding an ideal job based on their qualifications. Our team has decided to focus on employment websites where we intend to discover the significant qualifications that fulfill the expectations of employers and in turn enable job candidates to grasp the employment opportunity. Each of us will be scraping data from 4 different recruitment websites which include JobStreet, Indeed, Glassdoor, and LinkedIn to derive our individual understanding and analysis in preparation for modelling. The objective is to facilitate job candidates to find out what job opportunities are offered and recommend the most ideal job related to their skills and qualifications.

### INTRODUCTION / BACKGROUND INFORMATION

Job title is a comprehensive yet short and concise form description that conveys all the relevant information applicable to a job. The job title usually consists of professions, position, and the order of responsibility of any specified job. Substantial value is linked to job titles both internally within corporate organizations and to individual employees. Organizations arrange all employees in an organization chart on the essence of job titles. This has a relevance for issues like salary, level and degree of responsibility, employee selection, etc. Employees derive value from their individual job titles as a source of self-identification and this can have a major impact on their commitment and motivation. Categorizing job roles in accordance with the job information is a subjective human resources performance, however, which leads to bias and inconsistency.

It is the norm for many graduates to look for jobs and most of them may encounter challenges in the process of job hunting. Many companies use Applicant Tracking System (ATS) to organize job applications and eliminate applicants who do not meet the required qualifications. According to Zippia about job search statistics, approximately half of the online applicants are eliminated by ATS for the very reason that the resume did not cater to the requirements given in the job advertisement and missed the opportunity for recruiters to look at their resumes. *(Cite: Zippia. “15+ INCREDIBLE JOB SEARCH STATISTICS [2022]: WHAT JOB SEEKERS NEED TO KNOW”. Oct 6 2022,* [*https://www.zippia.com/advice/job-search-statistics/*](https://www.zippia.com/advice/job-search-statistics/)*)* Other challenges include lack of experience and education (for example, industry certificates) can lead to ATS removing resumes or trying to receive industry certificates. Therefore, developing a job classification model can help process the job descriptions more systematically and algorithmically such that the classification model generates a balanced and consistent result.

### BUSINESS SCENARIO

In recent years, the advancement of job seeking online creates opportunities for many employees to obtain useful and meaningful information for their jobs, and hence recruiters search for suitable candidates by advertising job description and qualifications for each job role. Nevertheless, with extensive information of job descriptions and qualifications, it is mandatory to classify such information into certain job roles. Not only does it allow employees to readily look for the most ideal job, but also suggest appropriate roles for employees. Therefore, our team proposed a classification algorithm (machine learning) to predict job roles based on the job seeker skills and qualification.

The machine learning (ML) task of predicting suitable roles from job descriptions is Text Classification. This ML algorithm receives the input as a post of job qualifications, then predicts the suitable roles for job qualifications as the outcome. To implement this task, we firstly scraped the data about job descriptions and job requirements from various online job finding websites in Singapore. After gathering the dataset, we incorporated the ML models for multi-label text classification on the dataset to make predicted jobs from the job description.

### BUSINESS OBJECTIVES IDENTIFIED

#### Business Objective

Our recommendation engine employs algorithms to suggest content to users based on their skill sets and qualifications. By analyzing data about the user behaviors, such as their abilities and expertise, the recommendation engine can provide personalized recommendations that are tailored to the user’s skill sets and qualifications.

#### Measurable Success Criteria

Accurately categorize job roles with approximately 70% precision.

### MAIN AND SUB TASKS IDENTIFIED

Main Task: Develop classification models using job requirements as classifiers to help in identifying the most suitable job.

Sub Task:

* Identify required data to scrape.
* To analyze the association between job roles and job descriptions, and the extent to which can be comprising of a job role using job details, more specifically the job qualifications.
* To develop a model and use text classification algorithm to make predictions.
* Evaluate classification models to derive the best outcome to predict results.
* Discover insights gathered from the 4 targeted job roles.

### DELEGATION OF TASKS / RESPONSIBILITIES

#### Determine Business Goals:

* Define the ideas and business goals.
* Researching to understand users’ point of view.

#### Data Collection:

* Scrape required data from reliable websites.
* Ensure collected data is comprehensive and retrieved correctly without any missing information and in compliance with the PDPA policy.
* Data collected have labels included.
* Ensure data is combined with every member without any missing values or information.

#### Data Understanding:

* Present the relevant information from the data using charts and tables to offer a deeper understanding of the data.
* Provide accurate and thorough descriptions of patterns and insights of data.

#### Data Preparation:

* Prepare data by doing comprehensive cleansing and preprocessing on dataset.
* Transforming the data to ensure high quality for modeling use.

#### Modeling:

* Adequate comparison of wide variety of models.
* Derive best performance model.

#### Evaluation:

* Report the performance of the model on the test set, including relevant evaluation metrics and any comparison to a baseline or previous model.
* Summarize the main findings of the project, including any limitations or areas for future work.
* Include links to the code and any relevant documentation for future reference and replication.

#### Presentation:

* Use clear and simple language to explain the technical details of the model, such as the architecture and algorithms used.
* Use visual aids to help explain key concepts and the model's performance, to make complex information more accessible and easier to understand.
* Provide specific examples of how the model is being used or how it can be used in the real world.

**TASK ALLOCATION**

Team:

* Each person in the group would be responsible for scraping approximately 700 job openings per job role.
* Totaling an approximate of 2500 jobs scraped per person.
* Finding other graphs to showcase the new insights we can find in relation to our topic using graphs to showcase the visualizations found in relation to the analysis between entities.
* Pre-processing techniques like stop words and normalizing words will be used in both model training data preparation and model prediction data preparation.
* Code models with detailed explanations of understandings and insights found in the implementation of the models and provide explanations of our thoughts and understanding along the way.

Zhang Xiang:

* Scrape data from [indeed (data analyst)](https://sg.indeed.com/jobs?q=Data+Analyst&l=&from=searchOnHP&vjk=9531469c02984a52).

[indeed (data scientist)](https://sg.indeed.com/jobs?q=data+scientist&l=&vjk=bfd29b6f90c29083).

[indeed (software engineer)](https://sg.indeed.com/jobs?q=software+engineer&l=&vjk=be9864881f8c41fb).

[indeed (business analyst)](https://sg.indeed.com/jobs?q=business+analyst&l=&vjk=9531469c02984a52).

Joanne:

* Scrape data from [jobstreet (data analyst)](https://www.jobstreet.com.sg/en/job-search/data-analyst-jobs/).

[jobstreet (data scientist)](https://www.jobstreet.com.sg/en/job-search/data-scientist-jobs/).

[jobstreet (software engineer)](https://www.jobstreet.com.sg/en/job-search/software-engineer-jobs/).

[jobstreet (business analyst)](https://www.jobstreet.com.sg/en/job-search/business-analyst-jobs/).

Ulysse:

* Scrape data from [glassdoor (data analyst)](https://www.glassdoor.sg/Job/data-analyst-jobs-SRCH_KO0,12.htm?suggestCount=0&suggestChosen=false&clickSource=searchBtn&typedKeyword=data%2520analyst&typedLocation=&context=Jobs&dropdown=0).

[glassdoor (data scientist)](https://www.glassdoor.sg/Job/data-scientist-jobs-SRCH_KO0,14.htm?suggestCount=0&suggestChosen=false&clickSource=searchBtn&typedKeyword=data%2520scientist&typedLocation=&context=Jobs&dropdown=0).

[glassdoor (software engineer)](https://www.glassdoor.sg/Job/software-engineer-jobs-SRCH_KO0,17.htm?suggestCount=0&suggestChosen=false&clickSource=searchBtn&typedKeyword=software%2520engineer&typedLocation=&context=Jobs&dropdown=0).

[glassdoor (business analyst)](https://www.glassdoor.sg/Job/business-analyst-jobs-SRCH_KO0,16.htm?suggestCount=0&suggestChosen=false&clickSource=searchBtn&typedKeyword=business%2520anal&typedLocation=&context=Jobs&dropdown=0).

Clara:

* Scrape data from [linkedin (data analyst)](https://www.linkedin.com/jobs/search?keywords=Data%20Analyst&location=singapore&geoId=&trk=public_jobs_jobs-search-bar_search-submit&position=1&pageNum=0).

[linkedin (data scientist)](https://www.linkedin.com/jobs/search?keywords=data%20scientist&location=Singapore&geoId=102454443&trk=public_jobs_jobs-search-bar_search-submit&position=1&pageNum=0).

[linkedin (software engineer)](https://www.linkedin.com/jobs/search?keywords=software%20engineer&location=Singapore&geoId=102454443&trk=public_jobs_jobs-search-bar_search-submit&position=1&pageNum=0).

[linkedin (business analyst)](https://www.linkedin.com/jobs/search?keywords=Business%20Analyst&location=Singapore&geoId=102454443&trk=public_jobs_jobs-search-bar_search-submit&position=1&pageNum=0).

### REFERENCE

##### Description

1. Install and import necessary libraries.
2. Create URL template for scraping the 4 job roles.
3. Create Dataframe for data scraped using Pandas: [‘Roles’, ‘Title’, ‘Company’, ‘Link’, ‘Description’].
4. Search for elements within the web page to scrape the information that will be stored into the Dataframe.
5. Remove duplicate rows based on conditions specific to each member’s.
   1. E.g. Using Link column
6. Create our individual csv files.

##### Scraping Website

* [https://sg.indeed.com/jobs](https://sg.indeed.com/)
* [https://www.jobstreet.com.sg](https://www.jobstreet.com.sg/)
* [https://www.glassdoor.sg](https://www.glassdoor.sg/index.htm)
* [https://www.linkedin.com/jobs](https://www.linkedin.com/)

##### Python Library

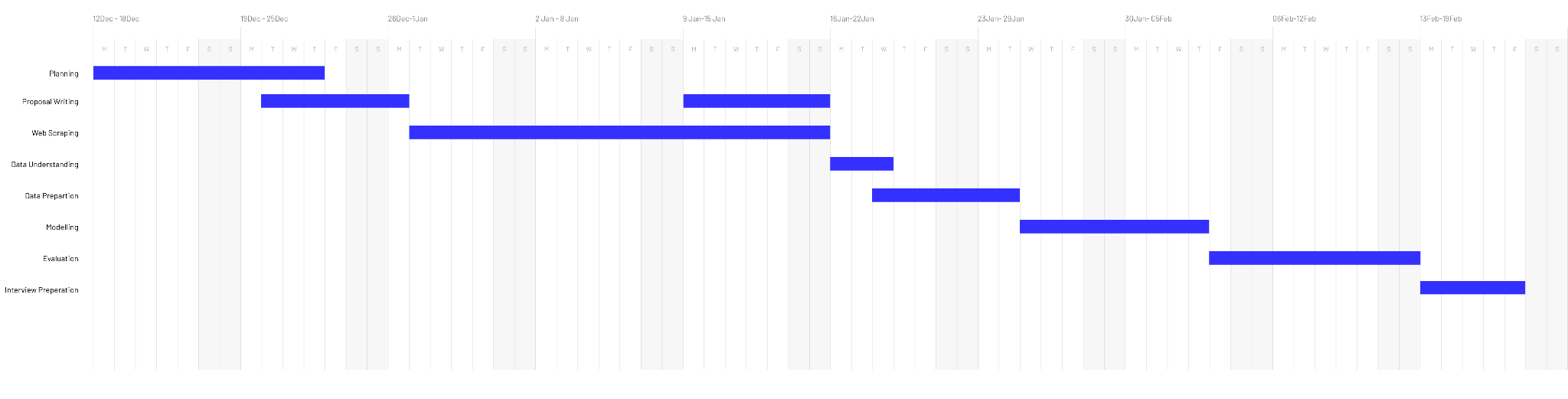
* BeautifulSoup
* Selenium
* Pandas
* Time
* Undetected chromedriver
* Math

##### Application

* Download chrome driver file for Chrome Driver:   
   <https://chromedriver.chromium.org/home>
* Steps to identify how to attain chrome header: <https://stackoverflow.com/questions/4423061/how-can-i-view-http-headers-in-google-chrome>
* Google Chrome User Agent for selenium:  
  chrome\_options.add\_argument("user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/108.0.0.0 Safari/537.36")  
  108.0.0.0 => Depends on the chrome version used.
* driver = webdriver.Chrome(executable\_path=r"C:\Users\User\Downloads\chromedriver\_win32/chromedriver" )

#### Timeline

##### Gantt Chart

****

##### Link

[**https://www.figma.com/file/1VWjQ4LpB9b8GK5nCm1R51/Text-%26-Social-Analytics-Gantt-Chart?node-id=0%3A1&t=0PNbjH2qoN8q3bST-1**](https://www.figma.com/file/1VWjQ4LpB9b8GK5nCm1R51/Text-%26-Social-Analytics-Gantt-Chart?node-id=0%3A1&t=0PNbjH2qoN8q3bST-1)