**Project Overview**:

The objective of the project is to determine the factors responsible for predicting the salary. The focus is on identifying the key words that distinguish the roles using Natural language processing where we tune the system to identify the key words that are important and goes with the title using various models and parameters.

The project is mainly split in to three parts:

1. The first part was to acquire the data with all required features necessary for prediction.
2. Once the data is acquired, I needed to explore in depth for insights for further analysis.
3. Once I know how my data looked like, my final step was to fit appropriate models and check which model is performing better and meeting the requirements.

**Process Overview**:

For this project I used X-path (XML path) to scrape Seek website, beautiful soup to parse the text and jupyter notebook for modelling and prediction. My focus was mainly on Data Scientist and Data Analyst roles with Sydney and Melbourne as my preferred locations. Using this search, I could able to scrape 7500 jobs. For each job, I obtained job title (Analyst, Manager, Data Scientist, etc.), type of work (contract / full-time/part-time), description, Industry (IT, Accounting, etc.) and salary as my features.

The challenging part of this project was exploratory data analysis. The data was raw and were to bring to uniformity for analysis. Firstly, I looked for duplicates in the data and removed them. Then looked for missing values in the data, and further filtering the roles that aren’t relevant for my analysis. After all the cleaning, I was left with 1250 observations for my analysis. To make it little easier, I tried binning various industries to few major industries.

For the first question, the goal is to predict salary. Surprisingly, there were only 35% of salaries listed and I must predict for the rest 65%. How do I do that? Firstly, the salaries were given in hourly, daily and annual format, I converted all of them to annual for uniformity. The only way I can do this is by grouping the job titles to junior, senior and manager and compare with the Industry (the pay differs from industry to industry based on the title) and then imputed the missing salaries with the median values.

**Modelling**:

1. The first model I chose was Linear Regression specifically using Lasso regression. The reason I chose this was to find out the top most important features in predicting the salary (topmost features I got were: **Work type**: Contract, **Industry**: Information Technology, **experience** **level**: senior, **title**: Data Manager). This model returned me R2 score of 0.30 which means 30% of variance within the salary is explained by the features.

I then treated this as classification by ranking the salaries. I tried both Logistic Regression and Random Forest. Both the models gave me better results; however Random Forest was giving me better score compared to Logistic Regression with the same list of top features that I got with Linear Regression.

1. For the second question, I used TF-IDF Vectorizer and simple stop words. TF-IDF which not only considers the number of times a word appears within a description but also compares it across all the descriptions. Stop words are used to remove the high frequency words that does not add value to the information. I fit 4 models (Logistic Regression, Random Forest, Naïve Bayes, SVM with linear and rbf) and attempted to predict the key words for the job titles (Business Analyst, Data Analyst, Data Scientist, Data Manager). I tried 4 different models with and without cross validation. The best predicted model for this was Logistic Regression with L1 penalty.
2. The top skills predicted for different job titles are:

* A Business Analyst is expected to have basic technical skills, most of the skills required are agile, functional, financial and planning.
* The Data Analyst should have SQL, tableau, Azure, Hadoop, AWS as top technical skills.
* The Data Manager depicts more of managerial skills like strategy, responsibility, budgeting etc. which every manager is expected to have.
* The keywords for Data Scientists do not exhibit a useful pattern due to lesser observations (only 38 observations) as well as vagueness in the job descriptions. Some of the relevant words which were picked up as top skills are AI, machine learning, modelling, predictive etc.

**Conclusion**:

The analysis was based on very few observations and it can always be improvised with the larger data and I would look upon working on it further by scraping in future for better analysis.