# Part 1 – Data Preparation and Pre-processing

**Dataset Description:**

The dataset is composed of data taken from SEEK job market and is composed of a CSV file just under 900MB large. Within the CSV file there are 12 columns (excluding the ID) which carry a range of metadata about the job listing. This information includes the company, date, classification, requirements, salary and the location of the job. The data ranges over the span from October 2018 until March 2019 with 318,477 entries.

There are 3 main components of the job which will be studied over the time period, this includes classification/subclassification, location and salary (lowest/highest). Therefore the relevant columns will be Date, Location, Classification, SubClassification, Lowest Salary and Highest Salary. Optional columns to extract information from include the Company, Area, and JobType. Due to the wide variation in formatting with columns such as Title, Requirements and FullDescription, these columns will be discarded along with ID.

**Dataset Preparation and Pre-processing:**