# Task 1

In an ever-competitive job market, finding employment can be quite difficult. Finding a suitable employment opportunity can become quite time consuming for job seekers with a medium sized posting having an average of 550 words [[source]](https://theundercoverrecruiter.com/tips-writing-job-descriptions/). On average, it takes approximately 15 applications to land a job interview, and around 10 interviews to secure a single job offer. This means that an applicant would need to apply for 150 job positions before receiving a single job offer [[source](https://talent.works/2017/09/22/how-long-does-it-take-to-get-a-job-60-days-if-youre-in-hr-or-sales/)]. If each job listing takes three minutes to read, a job seeker would need to spend 5 hours reading listings, plus additional time to complete the employer’s application process.

Currently, job seekers must scour through multiple sites worth of job postings and read through entire the entire listings in order to determine if they possess the required skill set whilst also ensuring the day-to-day tasks they will be undertaking are suitable for their career aspirations.

Job seekers would benefit significantly from a central repository of job listings, with each listing containing a summarization, that is restricted to a handful of sentences, and metadata tags of the required skills. This repository would streamline the job application process and enable applicants to apply for jobs at a notably faster rate by reducing the amount of time spent reading applications and filtering out listings that require skills the candidate does not possess.

A web-crawler could be employed to scrape through multiple sites that contain job listings, extracting the job title, job position and job description from each article. By utilising a web crawler that crawls multiple sites, the process of job listing collection can be automated and enable the central repository to contain a large quantity of employment opportunities.

By making use of Natural Language Processing tasks, the central repository can provide skill metadata tags and listing summarizations. For this use case, two tasks will be employed:

* Summarisation through the use of extraction-based summarisation NLP techniques
* Skill keyword extraction through the use of a Long Short Term Memory (LSTM) deep learning network and word embeddings.

# Task 2

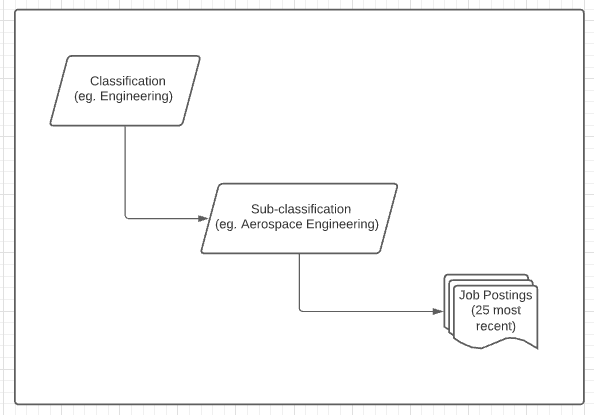
1. In total, two websites will be crawled. The first will Seek.com.au, a website which focuses on facilitating the match between jobseekers and employment opportunities and helping hirers find candidates for advertised roles [[source]](https://www.seek.com.au/about/). Seek has the ability to filter job listings according to the job classification, the job sub-classification, the job location, as well as the job work type (full time, part time, etc) and the pay.

The other website will be thebalancecareers.com, a site home to experts who provide clear, practical advice on job searching, resume writing, salary negotiations, and other career planning topics [[source]](https://www.thebalancecareers.com/about-us).

1. Seek.com.au was chosen as the primary site containing job postings to be extracted as it is the largest provider of this service in the Asia-Pacific region with exposure to over 2.9 billion people and relationships with over 1 million hirers.

Thebalancecareers.com was chosen to be crawled as it contains a plethora of job skills on its site, curated by a series of experts in the field that remain up to date on the current job trends and their associated requirements.

1. As this task is only a prototype, the job listing coverage is limited to only advertisements posted to SEEK, further restricted to the 25 most recent posts per sub-classification.



As of writing, Seek lists nearly 3,000 job posts for the Developers/Programmers sub classification of the parent Information & Communication Technology classification. By restricting the number of job listings per sub-classification to the 25 most recent, the resulting data set is of a much more manageable size. There also exists a plethora of other sites that contain job postings, including but not limited to: jobsearch.com.au, careerone.com.au, adzuna.com.au, indeed.com.au. These previously listed sites also restrict the job location to Australia. If one were to create a world-wide central repository, the crawler would need to crawl through an unknown but presumably large number of sites that exist for this purpose worldwide.

Balance Careers provides a large number of employment requirements from their handful of experts in the field, but these requirements they provide are purely a small sample of the total number of skills required. Due to the ever-evolving nature of careers and their required skill sets, choosing just a single site to represent all of the possible skills is severely restricting. Using multiple sites, or by employing industry experts that can stay up to date on current and emerging competencies.

1. The layout of seek.com.au is quite simple. For each query, a list of results is returned. The crawler the iterates over each result. This result is a single job posting which contains the job description which is extracted through the use of BS4. SEEK appears to employ one of two layouts for each job posting, one where the description exists inside a ‘div’ with the attribute ‘data-automation’ being set to ‘jobDescription’ otherwise the job description is just the next ‘div’ element after the job title.   
   The webpage ‘https://www.thebalancecareers.com/employment-skills-listed-by-job-2062389’ contains a list of professions with links to a page dedicated to each profession. These individual profession pages contain a list of skills stylised as an unordered list with each element having no class alongside paragraph content that proves to not be valuable to our case.
2. The robots.txt file present in the majority of websites contains the URLs that are not allowed to be crawled. In this case, both sites’ robot.txt files shows that all the URLs present on the site and are allowed to be crawled. To prevent copy right issues, the logos from the individual companies will not be stored.
3. The job description metadata will be supplemented with the skills extracted from thebalancecareers. The extracted skills will be used in conjunction with a manually labelled data set to train a RNN on what a skill is, which can then be utilised to extract skills from the job descriptions.
4. The SEEK web crawler will take the parent element that contains the job description and then find all children elements that contain text. The text from this list of elements is then joined and stored as the DESCRIPTION. The crawler for thebalancecareers will find every <li> element on the individual profession page with no class assigned to it. The text from each list item in the list of list elements will be appended to a skill list if that skill is not already present in the skill list
5. CODE
6. SCREENSHOTS
7. The pandas package will be used to construct a file with the list of jobs and their accompanying description. The file will be a csv file with the following headings:

|  |  |  |  |
| --- | --- | --- | --- |
| CLASSIFICATION | SUB-CLASSIFICATION | TITLE | DESCRIPTION |

‘DATA\_WRANGLING.PY’ The ‘’ file will be employed to clean the stored data. This is done through the re package in python, more specifically the sub function which substitutes a given regular expression with another supplied string. A series of substitutions are used to remove symbols present in the text including quotes, the newline character. Another series of substitutions are employed to remove certain words from the FIELD and TITLE columns.

# Task 3

## NLP 1 – Text summarizer

1. Lit review
2. Rationale for selection of NLP task

This task was chosen in order to extract the key sections of the job postings, allowing the prospective applicant to read a quick summarization of the post without needing to read the whole thing.

1. The input to the summarizer will be the job description, with the output being the 7 most important sentences as determined by the extractor. The

## NLP 2 – LSTM RNN Keyword Extraction

1. Lit review
2. Rationale for selection of NLP task

Job skills are the common link between job applications, applicant resumes and job listing by companies. Identifying skills in job postings is a significant problem and can provide a pathway for job seekers and hiring organisation. By ‘tagging’ each job listing with the required skills and enabling users to filter jobs by these skills would drastically improve the job search process.