# 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

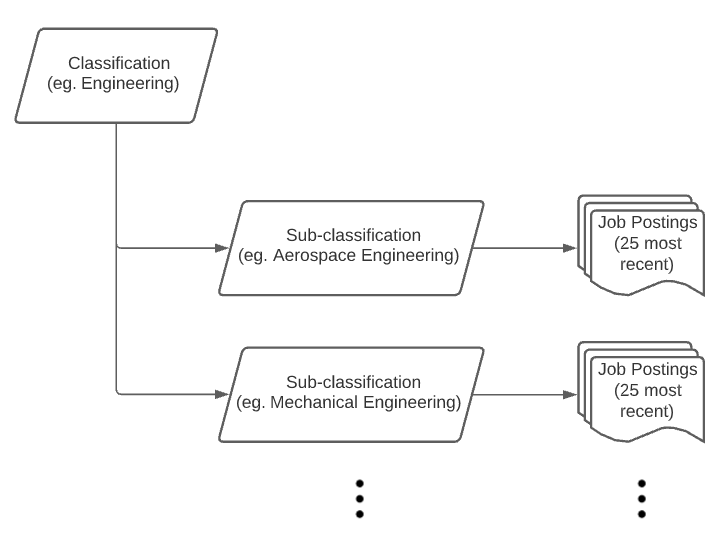
1. In total, two websites will be crawled. The first will Seek.com.au, a website which focuses on facilitating the match between jobseekers and employment opportunities and helping hirers find candidates for advertised roles [[source]](https://www.seek.com.au/about/). Seek has the ability to filter job 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 second site to be crawled will be thebalanacecareers.com. This website provides numerous articleson a wide range of topics ranging from finding a job and human resources to succeeding in the workplace. These articles are carefully curated by a select group of industry experts with experience in 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 relatively simple. The search filters are all stored in a html section tag which can easily be extracted through the use of the BeautifulSoup4 (BS4) python package. The selected filters are then used to generate a URL to show results that meet the filters’ criteria. The query results are returned in a paginated format with each page containing around 20 job listings. This pagination hurdle can be solved by appending the *?page=x* parameter to the query URL and incrementing the *x* until we extract the maximum allowed listings of 25. A loop can then be employed to iterate over each article on each page and extract the link to the full listing. Once on the single listing’s page, we aim to extract the job description which can be located in three different elements depending on the layout of the listing. The description can be in a div element with the ‘data-automation’ attribute set to either ‘jobDescription’ or ‘jobAdDetails’. If the description can’t be found in either of those elements, it will be located in the next div element after the job title.

Only a single page from thebalancecareers will be crawled, <https://www.thebalancecareers.com/employment-skills-listed-by-job-2062389>. This web page contains a plethora of possible job titles, divided by their industry sector, with each job title actually being a link to a dedicated page for this profession. Each individual professions’ page is laid out differently to each other, but each share the fact that they all contain multiple dot point lists of skills. These skill lists are stylised as unordered lists with each list element having no class. There also exists paragraphs of content elaborating on these skill lists, but these paragraphs prove to be of no value for the task at hand. By employing BeautifulSoup once again, the content of each skill list can be extracted and subsequently written to a file.

1. 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.
2. The job descriptions metadata will be supplemented with the title, classification and sub-classification. The complete skill list, created by appending each individual professions’ list of require skills obtained from thebalancecareers, will be used in conjunction with a manually labelled data set to train a Recurrent Neural Network (RNN) on what a skill is. This RNN will then be used to extract all the skills contained in the job descriptions.
3. The SEEK web crawler will take the parent job description element and combine all the children p tags into a single string. This string will be stored in a csv file under the heading DESCRIPTION. The job title can be obtained by casting the HTML pages’ title attribute to a string and then storing it in the csv file under the TITLE heading. The classification and sub-classification will be available as variables in the workspace but can also be obtained from the webpages’ URL.

The crawler for thebalancecareers will append the skills found from each individual profession page to a single skill list if the individual skill is not already present in the global skills list. This skill list will be written to a csv file with the skills being under the TEXT heading. An additional column titled TARGET will also be written the value of ‘1’ for each skill. This TARGET column will be utilised when training the RNN.

1. CODE
2. SCREENSHOTS
3. 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.

1. Summary and visualisation of the harvested data. Preliminary exploratory analysis

# 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.