# Task 1

Overarching issue: extracting required skills from job posting for job seekers

Webcrawler: webcrawler targets domains with job postings

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 450 words [[source]](https://www.linkedin.com/business/talent/blog/talent-acquisition/stats-that-will-change-the-way-you-write-job-posts). A job seeker must read every job posting to find whether they have the required skillset and whether the day-to-day tasks they will be completing are something they are interested in.

A web crawler will be able to target domains that provide job post hosting and extract the textual content from these websites of interest.

The first NLP task will be the summarisation of the job posting through the use of extraction-based summarisation.

The second NLP task will be skill keyword extraction through the use of LSTM and word embeddings, combined with a 2000 row manually labelled training set. The training set is comprised noun-phrases extracted from the first job posting from each sub-classification.

# 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/). 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).
2. 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 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.
3. 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. There is a plethora of Australian sites that contain job postings, including but not limited to: Facebook, jobsearch.com.au, careerone.com.au, adzuna.com.au. These listed sites are also all restricted to Australia. The SEEK group deals with job postings from all over the APAC region, with an unknown but presumably large number of sites existing 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.
4. 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.
5. 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.
6. 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.
7. 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
8. CODE
9. SCREENSHOTS
10. 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