**Research Question**

Identify the 5 most prevalent characteristics sought after (or found) in the hiring process for one of the Data Science roles (Data Scientist, Data Engineer, or Data Analyst)

**Subquestions**

* Identify skills that employers are generally looking for in the Data Science role?
* Identify common sought-after work conditions by employees in the Data Science role.
* Identify common personality traits of employees in the Data Science role?
* Identify common personality traits that employers seek in the Data Science role?
* Identify common skills or experiences of employees in the Data Science role?

**Process of Communication**

Communication between the client and knowledge engineer groups was at first done through Teams meetings, in which each group clearly indicated their research question and took some times to work up an understanding of the topic. Due to some difficulty caused by one of the provided sources being inaccessible, some back and forth communication was required in order to decide how to handle this dataset being replaced. In the end it was decided that the smaller replacement dataset could be used, however it may not be required for the answering of the subquestions.

During the intermediate progress meeting the knowledge engineer group indicated that the workload of the provided 5 sub questions would be too big for the duration of the project and some discussion took place to decide how to adjust the question accordingly. Based on the engineers’ suggestion and some clarifying of goals and applications the initial questions were adapted, allowing the client to focus on combining the skill defining terms from the provided datasets and creating a tool to compute the number of occurrences of these terms as well as their links, essentially creating a tool that can be used to find commonly shared terms and combinations of terms amongst the provided job description datasets. As this aligns with some of the purposes of the original questions, this proposal was accepted. Along with the proposal of the tool, the focus was put on answering the first and fourth sub-question originally provided.

Communication near the end of the project was rather sparse, however due to a direct connection between the groups in which questions could be asked at any time, this did not seem to be a problem.

**Quality of Results**The engineers used a predefined set of skills to search through the job postings and find the absolute and relative occurrences of the skills in the postings. The predefined set of skills was obtained through filtering through the provided articles and extracting terms that were suitable for describing a (set of) skills. The final ranking of the skills show both soft and technical skills intertwined as agreed upon with us. The top skills were machine learning, technical, management, statistics, and analytical. A Knowledge Graph of the co-occurrences of skills is visualized, where skills are nodes and edges are co-occurrences. The size of a node indicates the number of co-occurrences. Based on the selected attributes, the most co-occurring attributes will remain in the graph, which could give an indication on whether they are useful.

The final results answer the question which of the skills, gathered from the provided articles, are most represented in the provided job descriptions. Thus, we could consider the team’s work as successful in answering our main question.

It should be pointed out here that the sub questions 1 and 4 that were focussed on after being asked to narrow down the focus were directed at finding the hard and soft skills sought after. Although the presented results do not offer a strict distinction between the two types of skills, which was not able to be conveniently categorised, as the current ranking offers a combination of both hard and soft skills, this was not found to be an issue.

**Presentation of Results**

The results were presented using a user interface, where the predefined skills can be selected. By selecting skills, their total and relative occurrences are shown in a table. The interface also visualises the knowledge graph, where the number of co-occurrences can be controlled.

Additionally, it is possible to filter both representations of the results. In the case of the ranked terms, it is possible to add specific terms, either selected from the list of terms created using the provided articles, or through the input of an entirely new term. Resulting from this is then the list of default terms combined with the added terms to form a new ranking.

In the case of the knowledge graph, it is possible to filter the graph based on the number of co-occurrences between terms, allowing the user to specify a minimum number of co-occurrences required in order for a term to appear in the graph. This in turn lets the user find more specifically which combinations of terms are present at least the specified number of times.

**Future Possibilities**

Some extensions that could be considered in the future would be to extend the tool using a concrete list of soft and hard skills, or a measure by which terms can be categorised such that a clear distinction between the two type of traits can be used to filter through the job applications, offering a way to let the user find an even more fine-grained list of terms for either type of skill. A similar extension could be made to the knowledge graph representation such that it is possible to distinguish more easily between what type of skill is part of the co-occurrences displayed.

It may also be a possible extension to introduce Natural Language Processing (NLP) to help understand and gather additional terms and descriptions for skills. Additionally, NLP could be used to find synonyms and terms with slightly different spellings between languages or dialects (or plainly terms being written wrongly). NLP could finally also be used to help translate between languages, to extend the reach of the tool to applications from all sorts of languages. This may be particularly useful to help international students or multi-lingual job applicants.

Additionally, the tool could be extended to also answer the main question for the other data science roles instead of just one. The subquestions that were not considered in the current project could also be included in the future.

Finally, it may be a consideration to extend the use cases of the tool with non-data science jobs and roles. By doing so, the tool could potentially be extended such that it allows for the comparison between jobs based on their most common sought after skills.