Capstone Project proposal

A good data scientiest should be good in asking interesting questions.

So the question that lead me to the formulation of the idea for my capstone project is : what programming languages, IT skills in general, soft skills and experience should we master for getting a job in a company like (say) Google or Amazon?

Thus the problem I want to consider is how to find the perfect match between the skills, experience and more in general the professional profile of a person and a given open position in a company's careers site.

On the one hand, I believe that solving this problem could be useful to many people trying to switch careers and/or job seekers.
On the other, the methodologies developed during my capstone project challenge, could be useful to reqruiters too, as they might need a first automatic screeing of the hundreds of CVs they receive everyday.

For solving this problem, I have chosen a text dataset that was created starting from the Google's careers site which shows the open positions at Google https://careers.google.com/ In addition, I will integrate to this dataset also other data scrapped from the Linkedin API and other sources in order to complement the initial dataset.

The idea is to match those open positions with the skills of a particular candidate or job seeker that is evaluating the aforementioned opportunities. The dataset is mainly about unstructured data, in this case bulky text.

Thus, the analysis has to be modeled as a NPL(Natural Language Processing) problem and as a matching problem in which I will try to find the best match between a candidate and the available positions.

Describing briefly the steps of my work :

1- I will develop algothims to first analyze the text dataset for extrating possible patterns. Those patterns might be skills that tend to compare together for the same position and or historical data on what company hired what type of professional.

2- I will then use those patterns to create the best possible matchings and then

3- test the results with ground-truth data. I plan to use Python and different machine learning libraries in particular the ones for NLP.

As for the deliverables, I have created the following github repository
https://github.com/alketcecaj12/CapStoneProjectIntermediateDataSceinceInPython.git

There I plan to put all the pieces of my capstone project, i.e. the code, the datasets and the related documentation.