TIS Project Progress Report (Job Recommendation System)

1) Which tasks have been completed?

We have taken a public dataset as reference and extracted the job description column from the dataset. Similarly, we have implemented document parsing techniques in order to parse every page of the resume that gets used. Next, we have implemented the a function which goes through every job description in the dataset as well as every page in resume, does preprocessing on it like cleaning, lowering the case, removing stopwords, taking unique words and then matches it with our global skill set dictionary in order to return the extracted keywords for the job description as well as resume respectively. Now we have our 2 sets of keywords (one set from description and one set from resume), we tried different similarity measures like jaccard similarity and cosine similarity and decided on using cosine similarity. The reason for choosing cosine similarity was because we want to give preference for duplication since we might see instances of the same skill at multiple occurrences in either the resume or the job_description and we wanted to use that as a criteria while ranking jobs.

2) Which tasks are pending?

Currently, we have used a reference public dataset in order to extract keywords as well perform similarity measures on it, however we are still in the process of finalizing our combined dataset by combining public dataset and web scraped data which will best suit the requirements of our project . We are also yet to create an interactive user interface for our project by designing the front end with React and Flask.

3) Are you facing any challenges?

Currently, One major challenge that we are facing is to retrieve jobs data in real time. We are relying on web scraping LinkedIn to collect jobs data whereas the ideal way to fetch data in real time would be using API. All the API's we explored require either production level access or just provide limited response for development use cases like ours. However, jobs data is relevant for a long period, so even though the jobs we would recommend to users won't be based on real time data, it will still be relevant for them, since we plan to run our scrapper regularly throughout the day or a week.