Data Acquisition and Analysis

Duration : 2 weeks

Given the provided dataset was a bit messy and the following pre-processing would be difficult, I have written a snippet of code to fetch predefined scope of job postings, namely from financial institutes in London, to acquire a easier-to-use datasets from indeed.com. The code fetched location, company name, job title, a brief job summary on the meta page and a full description people will see when they click into the subsequent posting page. The code file is uploaded to the main GitHub Repository.

The scrapped data which was stored as pickle file is then retrieved for pre-processing to take place. The pre-processing includes: punctuation removal, contraction expansion. tokenized, stop-words removal and stemming and lemmatizing. Some qualitative analysis were also included in the EDA file. Then for occupations with high frequencies, a word cloud is generated to visualise the hottest words. However deeper insights should be obtained after further meeting.

Embeddings and Language Representation models

Duration : 2 weeks

As understanding in NLP is deepened with the course, I found it necessary to also enrich the technical aspect when designing, implementing and deploying and up-scaling the actual NLP pipeline for skill identification. Therefore time was spent into investigating the existing works in embedding as well as different language models. Amongst different works, I found it would be promising to utilise transformer models which ignores the sequence but have position embedding in place. As illustrated in the work of Yu Li, the application of this model combined with other modules including keyword matching etc. achieves higher accuracy as well as 20 times faster prediction rate comparing with ordinary BERT model. Sample works on using PyTorch and Spark NLP are also explored.

Case study on Skills-ML

Duration : 3 weeks

It is crucial to understand the skeleton the alternative approaches and amongst the existing works skills-ml provides the most detailed documentation as well as source codes.

Next step: Test on Skills-ML

Duration : 1 weeks

To establish a benchmark for our future work, I am planning to using our data from indeed.com on the skills-ml pipeline and using this sample pipeline to identify potential spots to be improved.