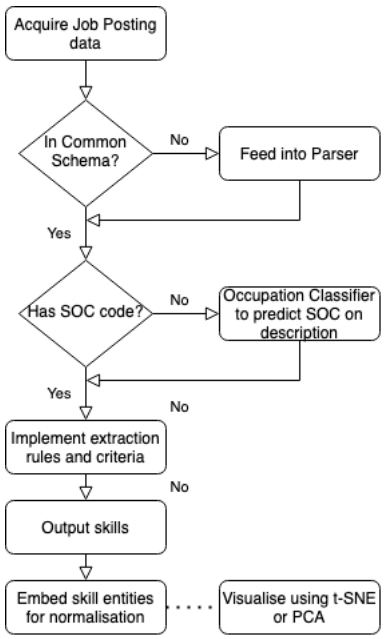
Application on Skill Extraction: Project Handover

## Overview

This project is oriented around the skills requirement aspect of job postings on the online recruitment platforms. The coding contains sections specializing and sequencing on the following tasks: acquire job posting data through scraping indeed.com by specifying predefined search criteria, parsing the scraped textual into job Common Schema, predict the occupation classifier based on texts of job descriptions through means of doc2vec embeddings (which is pre-trained based on previously concatenated dataset), extracting skill entities disambiguated using occupation classification coded (SOC), embedding the previously extracted into normalized vectors that could be then visualized through either t-SNE or PCA to showcase the overlapping and differences. The formulated pipeline for processing of textual skill entities provides a ready-to-use tool for researchers in labour markets to address problems such as evolution of job titles, discrepancies of skill requirements between different industries, lifecycle of a job inside an organization and identify what crucial skills are needed for promotions and sustainability etc.




## Instructions

The pipeline utilized the functionalities of open-source library Skills-ML to perform the extractions of skill entities with several modifications. The general instruction for usage of this package is demonstrated on the 'Skill-ml tour' notebook file.

The main part of the code is modified into the following notebooks with comments on directions to use, with index representing order to execute. The folder main folder contains three subfolders, code, data and output, each functioning as a referencing path to store data file as well as output charts. The file 04a performs a basic a basic demonstration on addressing the research problem addressed in the vignette 1 in the working paper. In addition, in the requirement folder contains the exported conda environment file.