**Job Matching System using Natural Language Processing (NLP)**



**Context and Problem:** Identifying candidates from underrepresented groups, regions or areas are critical for improving diversity and inclusiveness in the hiring process. In recent times, the use of AI screening systems allows recruiters to control and avoid human bias. The use of advanced NLP and machine learning algorithms provides more insightful analysis and understanding of both the talent/candidate (experience, education, skills, career trajectory, etc.) and the employer’s job requirements to provide the most comprehensive screening for potential mutual fit. The algorithm uses Sentence Transformers from NLP to understand CFW (cash-for-work)1 & CFIA (cash-for-internship apprenticeship) from the participants’ profile information, including educational background, training, work experience and preferences.

**Database:** The data used come from the partner’s SharePoint which is connected to a power bi Dashboard for automatic visuals updates. The results got from the system are stored in that SharePoint

**Repository: the documented notebook**

<https://colab.research.google.com/drive/1Im4q45mVpdasgR6reg8ddpo1ZzPWlBEB#scrollTo=9zLhHTft7gM_>

**Process**:

The system requires 3 files to be able to run:

* Participant’s data
* Employer’s data
* Mapping data

The Educations\_specialization variable (when it exists) or previous\_work\_experience (1, 2, when it exists) or preference\_work (1, 2, when it exists) from participant data is mapped to a position using the mapping data. The mapped position is then the one used for the matching.

In the first step (semantic matching) the algorithm uses all-mpnet-base-v2 transformer model the mapped position from participants and job positions from employers to a 768-dimensional dense vector space and assess similarity scores between participants vectors and employer vectors in that space.

In the second step it filters potential matched candidates as per employers' requirements (disability, Gender, Age, region, maximum number of participants). The system returns then the participants that match the exact needs of the employers. It also helps to assess the fairness (region and sex) of the process.

The pipeline is built such that the algorithm complexity is of order of linear combination of inputs sizes. In that way, the system can handle billions-sized data in few minutes.