Discovery of validation against skills-ml

1. The overall accuracy, calculated from the ratio between the number of correct skills detected and the total number of skills identified, is 0.66 for finance sector and 0.75 for telecommunication sector.

2. By comparison between the skills extracted using skills-ml and the manual label entities, there is an obvious lack of finance-related skills. This drags back the overall performance significantly and the reason can be attributed to the fact that people speak a specific 'language' in the finance sector. However, this issue can potentially be addressed by training the embedding model using our dataset from indeed (which would require access to the major dataset) and also enrich the finance part of the ontology if possible.

3. The accuracy of skills-ml is not consistent across job postings with different formats (layouts), identified from the fact that the same skill entities are extracted from some postings but not all that contains them. This could be addressed by cleaning the data with the purpose to normalise particular sections, namely 'Responsibilities' and 'Requirements'.

4. Word sense ambiguity is not properly addressed due to the lack of soc code, which is a highlight of skills-ml. However, this could be addressed by reverse engineering the soc code from search criteria and job titles, which by its nature reveals the category of the job.

5. (ps.) The quality of the job posting is not consistent neither however, as a small portion of postings does not contain any information regarding the skills requirements at all.

Conclusion

In conclusion, at this stage the skills-ml pipeline shows to some extent good potentials in skills extraction and the following improvements can be implemented:

1. Embedding model training on finance sectors to enrich the finance-specific skill entities.

2. Format normalisation of the document which could be useful in either choice of paths for this project.

3. Reverse engineering of soc code to empower the attribute of ontology in skills-ml.