Documentation of PoC 1 code

The entire PoC consists of a mix of R and Python scripts.

The PoC on a whole is not completely automated, but the different scripts were executed manually.   
The automated parts will be deployed on the cloud platform, the data gathering part will not.

PoC 1: consists of the following scripts:

1. Scraper scripts in R (not cloud),
2. Python notebook NLP (not yet cloud),
3. R script for matches : join all data, graph compatibility (cloud),
4. Shiny app : visualise graph (cloud).

Data gathering consisted of existing data (ESCO dataset), obtained data (University Antwerpen) and scraping data (universities UCL, ULiege, VUB)

# Scraper scripts in R

Three scraping scripts were written by PwC:

* *scrapeVUB\_V0\_1.R,*
* *scrapeULiege\_V0\_1.R,*
* *scrapeUCL\_V1\_0.R.*

Each of the scripts is similar in it’s logic, but is specific for each university. The main reason is that the HTML structure and links to courses is different each time.

The most important information it scrapes is:

* The name of the course,
* The name of the education program,
* Description of the course content.

Some universities provide additional course information (such as learning outcomes or skills) aside from the course description. Where available, it was also scraped.

These scripts will not be automated, as they only were intended to gather the data.

It is possible to imagine building more general scrapers that allow scraping of different universities, or asking universities to provide the data themselves.

# Python notebook NLP (Not yet cloud)

2. NLP\_skill\_extraction.ipynb

The second script was made in python.

The script is documented in detail, so we’ll give a summary here.

It starts with reading the input data :

* ESCO dataset,
* Education programs, courses, course descriptions: from the universities.

The ESCO dataset contains a large list of 13.000 skills with a skill description attached. Each of these skills are linked to a list of professions/occupations in a graph database manner.

By using the course descriptions and the skill descriptions at hand, we can calculate the similarity between each combination of these two lists of text.

Then, we compute the cosine similarity between the vectors that represent the two lists of text.

This gives a value between 0 and 1, representing how close in meaning/context match.

We consider a threshold of 0.78, meaning that we do not associate skills to courses if they fall under that limit. The choice of 0.78 was not completely arbitrary, but came from an early evaluation of initial results and inspection of what was actually matching. Adopting this face validity, a score around 0.75 and 0.80 became apparent as possible cut of.   
Of course, it is always possible to revise this cutoff score.

# R script for matches : join all data, graph compatibility (cloud)

# Shiny app : visualise graph (cloud)