Data Engineering Track - Slot 1

The three slots in the Data Engineering track are focused on exposing you to the challenges of real-world data processing and analytics.

The overall track is designed to foster the following insights:

Insights:

- Difficulty of reliable data processing
- Limitations of real-world data structures
- Best practices to data management

In terms of concrete tasks you will be practicing the following skills:

Practice:

Acquiring new skills:

- Scraping
- Data pipelines
- DAG execution

Real-world Data Transformations and Handling.

- Missing Data Handling
- Data Selection
- Data Extraction
- Data Modelling

Data Engineering Slot 1: Job Crawling

For the first slot your task consists of implementing a tool that will allow us to automatically retrieve information from the job web-sites.

Crawling Target

The web-sites to be targeted are the following:

- jobs.ch
- indeed.ch

Feel free to also target other job portals of your choice.

Scraping Framework

You will use a framework for the collection of data from public web pages that is supplied by the BSc Data Engineering.

You will get the credentials before the initial introductory session for P&T 2.

In case you have not received your credentials please reach out to erik.graf@bfh.ch

We advise you to read through the initial getting started section and the tutorial in your language of choice.

- https://www.scrapingbee.com/documentation/#getting-started
- https://www.scrapingbee.com/tutorials/getting-started-with-scrapingbees-python-sdk/

The request-builder can be very useful to try different scraping strategies.

- https://app.scrapingbee.com/request-builder

Task Definition

The task for slot 1 is defined as follows:

1. Get familiar with the scrapingBee tool by consulting the tutorials

2. Analyse the structure of the web pages and develop a strategy that aims at

    - choose on of the two web sites listed above
    - download as many job posting texts for Switzerland for your chosen web site (at least 2000).
    - store the downloaded web sites locally, just store the html-pages and do not yet clean or analyze the data as this will be part of the task for slot2

Getting to the actual information you are interested in often means you have to first find a way to collect all links for those pages.

Finding those links can require different strategies to mitigate potential limitations on the number of results shown by webpages.

You should document those strategies and findings in the short report mentioned in the next section.

Deliverables:

The deliverables for this step consist of:

- Your code for the analysis as a Jupyter Notebook which contains the result of your last execution.
- Explain your thoughts in Markdown cells in the notebook.
- 2-3 page document (as pdf) or slides (as pdf) where you discuss the challenges, the limitations you encountered and the approach you devised to overcome them (you may also append this discussion in your Jupyter Notebook)
- a Zipped file contained downloaded samples (approx. 100 job description samples).
- Upload to a newly created repository for this track on BFH Gitlab and provide clickable links to these resources in Moodle - provide reporter rights to the grading lecturer (see Moodle)

If you have questions or require support please use the following link to book appointments:

https://outlook.office365.com/book/BFHBookingsbsc_data_engineering@bfh.ch/?ae=true