Data Engineering Slot 1: Job Crawling

Author: Michael Peter Schmid
Version 1.0
Last updated: 14.10.2025

Website Structure

I decided to use jobs.ch since in my opinion jobs.ch structure is easier to understand than indeed.ch. After exploring and navigating on jobs.ch as well as inspecting the page html code, I document my important findings:

1. As a first step I visited https://www.jobs.ch/de/ and then I scrolled down to "Jobs nach Berufsgruppen" This way I can filter for the jobs field I'm interested in.

**Eine weitere Möglichkeit, deine Stellensuche zu beginnen**

Entdecke deine Möglichkeiten anhand von Berufsfeldern oder wähle deine bevorzugte Region.

Jobs nach Berufsgruppen   Jobs nach Regionen   Meistgesucht   Top Städte

Administration / HR / Consulting / CEO	7805	Elektronik / Technik / Uhren	2407	Marketing / Kommunikation / Redaktion	745
Banking / Versicherungswesen	2494	Fahrzeuge / Handwerk / Lager /...	3920	Maschinen- / Anlagenbau / Produktion	3961
Bau / Architektur / Engineering	6589	Finanzen / Treuhand / Immobilien	2336	Medizin / Pflege / Therapie	6874
Bewachung / Polizei / Zoll / Rettung	396	Gastronomie / Lebensmittel /...	2769	Sport / Wellness / Kultur	291
Chemie / Pharma / Biotechnologie	756	Grafik / Typographie / Druck	101	Verkauf / Kundendienst / Innendienst	4998
Einkauf / Logistik / Handel	1003	Informatik / Telekommunikation	2932	Verwaltung / Bildung / Soziales	2450

2. Now when clicking on a link for example:
https://www.jobs.ch/de/stellenangebote/informatik-telekommunikation/
One will be directed to a site, where jobs in that field are displayed in the sidebar. And there are 100 subpages. It is possible to navigate trough these subpage by appending ?page={number}   where number is a between 0 and 100.
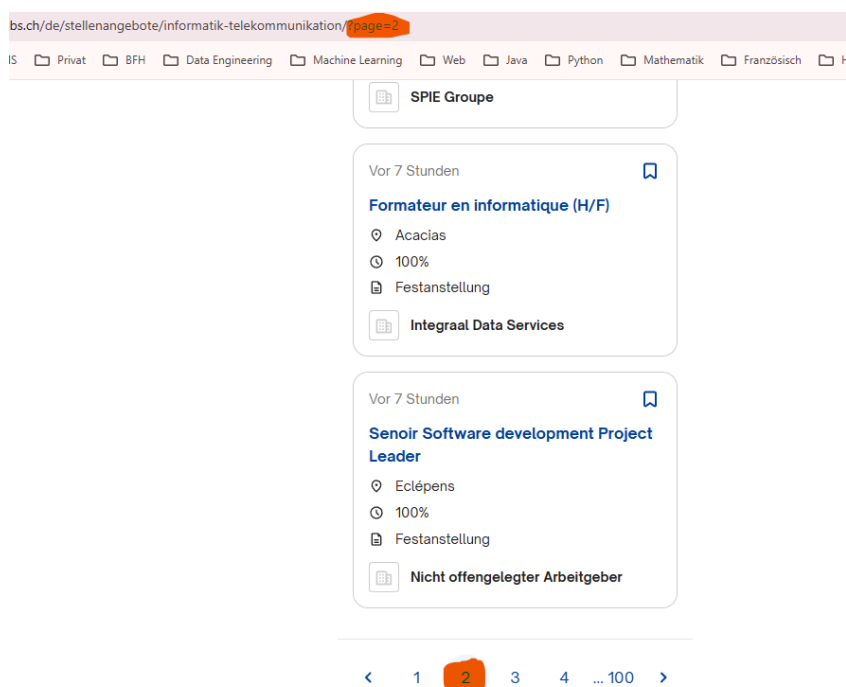
3. Now if we inspect the source code at this point we see that there is a div for each "job box" and within this div tag is an anker tag which contains the specific job id: So for each page there are multiple job id's inside these divs.

Strategy

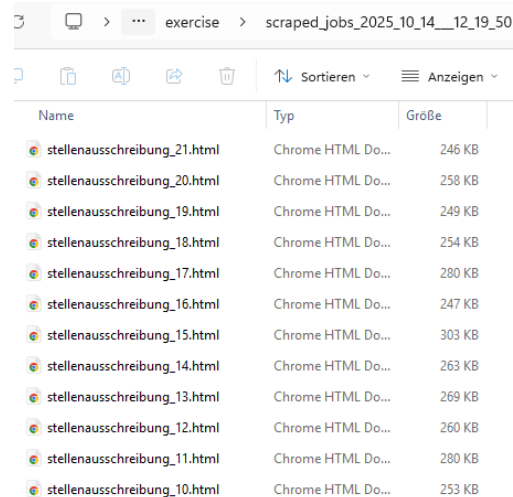Based on what I described in the previous chapter, here's my strategy for this task:

1. Pick some "Berufsgruppen" that I'm interested in. (in my case: "informatik-telekommunikation", "finanzen-treuhand-immobilien")
2. Collect Jobs for each "Berufsgruppe" for each page by calling https://www.jobs.ch/de/stellenangebote/{Berufsgruppe}/?page={number} Return the result for each page (number) and collect all the job-ids
3. Call https://www.jobs.ch/de/stellenangebote/detail/{job_id} / for each previously collected job id: For each result make a new file and write the result in it.

Limitations

- One script-run takes over 60min
- Only applicable to jobs.ch
- Tight coupling: Script will fail, if jobs.ch teams decide to change it's website structure or even renaming job field categories or other variable names in the URL.

Results

- A directory with job html pages:

- If one page is opened one can scroll down and will find the job description example:

(WoodTec group) explores sustainable wood processing and innovative applications of wood materials to support a circular bioeconomy. We are seeking a highly motivated postdoctoral researcher to develop and apply machine-learning and data-science methods, with a focus on computer vision, to enhance wood utilization. The successful candidate will play a significant role the EU-funded TIMBERHAUS project (www.timberhaus.eu).

Your tasks

- Develop machine learning models and computer vision algorithms for wood property analysis, sorting, and grading.
- Contribute to the TIMBERHAUS project by applying advanced image processing and AI methods for timber quality assessment.
- Integrate experimental data with computational models and collaborate closely with lab teams.
- Publish research findings, present at conferences, and support PhD students.
- Help shape new research proposals and assist in securing funding.

Your profile

- PhD in Computer Science, Data Science, Machine Learning, or a related discipline.
- Proven experience in computer vision (e.g. image processing, deep learning, object detection, segmentation) and machine learning techniques.
- Programming skills in Python and relevant ML frameworks (TensorFlow, PyTorch), experience with version control systems (e.g., Git).
- Strong publication record in reputable journals and conferences.
- Interest in wood technology or biomaterials is a plus.
- Excellent communication skills in English; proficiency in German is beneficial.

Our offer

We offer a fully funded 1-year position, with the possibility of extension, in an inspiring, multidisciplinary and international research environment at a leading Swiss institute. The successful candidate will have access to state-of-the-art infrastructure and opportunities for collaboration within the ETH Domain. Professional development, networking and participation in innovative projects are supported. We provide competitive salary and employment conditions in the Zurich region.
We live a culture of inclusion and respect. We welcome all people who are interested in innovative, sustainable and meaningful activities - that's what counts.
We look forward to receiving your complete online application including a letter of motivation, CV, certificates, diplomas and contact details of two reference persons. Please submit these exclusively via our job portal. Applications by e-mail and by post will not be considered.
Patricia Nitzsche,
Stv. Leiterin Human Resources / Dep. Head Human Resources

Questions?

Dr Mark Schubert Write an email

Group leader WoodTec group
Cellulose & Wood Materials

www.empa.ch/web/s302

Your future place of work

Empa
Ueberlandstrasse 129
8600 Dübendorf

Empa as an employer

Innovative, sustainable, meaningful activities
Creating added value for society
International, multicultural working environment
Freedom to create and develop
Culture of inclusion and respect
Excellent balance between different areas of life
Multiple award-winning and certified employer
Benefits for rail, mobile, childcare, catering, etc.

Good to know

Empa website