Take-Home Assignment

Your task is to re-implement a simplified version of one component of our infrastructure. Currently, we scrape data from multiple sources, integrate and transform it, and load it into our cloud database.

For this assignment, you will focus on the following:

Main Task

Scrape job data from the following website (using any programming language and libraries of your choice):

https://www.arbeitsagentur.de/jobsuche/suche?angebotsart=4&ausbildungsart=0&arbeitszeit=vz&branche=22;1;2;9;3;5;7;10;11;16;12;21;26;15;17;19;20;8;23;29&veroeffentlichtseit=7&sort=veroeffdatum

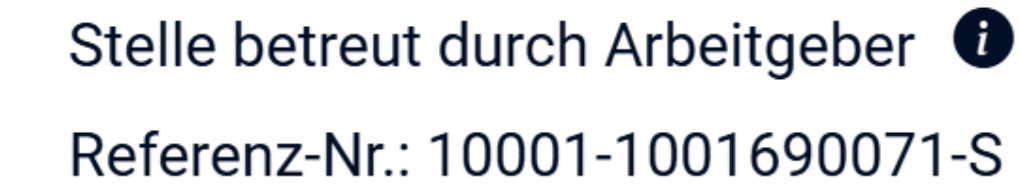
The scraped data should be structured in the following format:

| Profession | Salary | Company Name | Location | Start Date | Telephone | Email | Job Description | Ref.-Nr. | External Link | Application Link |

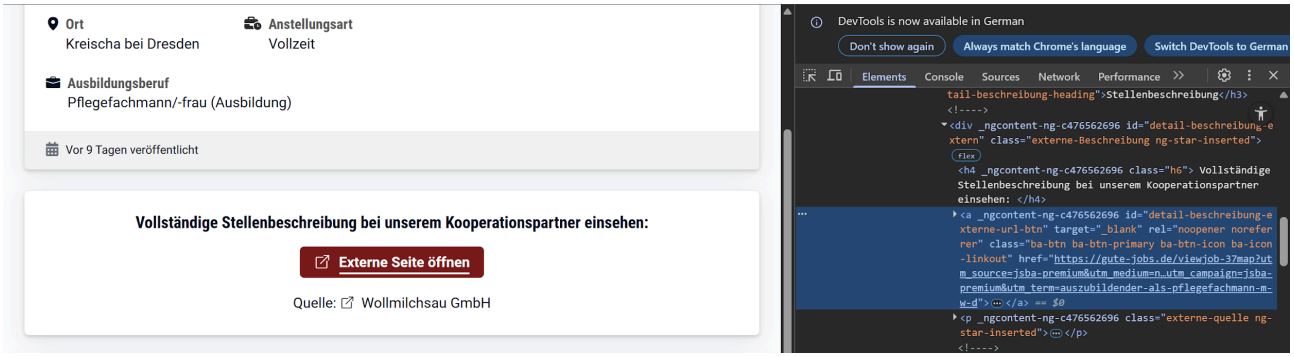
Field Notes & Examples:

Start Date: e.g., Beginn ab 01.08.2026

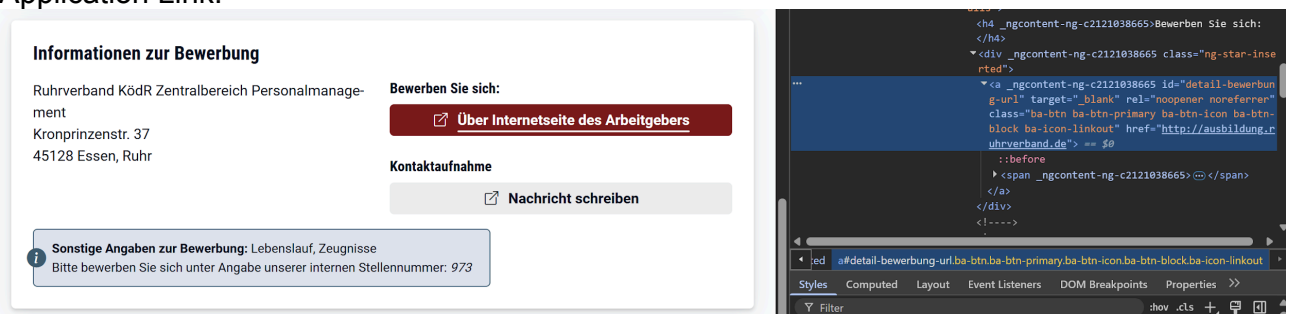
Ref.-Nr.:



External Link:



Application Link:



- Some jobs may have missing fields (e.g., email). This is acceptable.
- Some rows may contain only an Application Link or only an External Link; both cases are fine.
- You may encounter a CAPTCHA. You have to solve it the first time to access the full data and can choose to solve it more times to gather additional sample data.

Bonus Tasks (Optional)

To stand out, you may attempt the following challenges:

1. Transform & Load Data into a Database (Easy)

- Spin up a database (locally or via Docker).
- Create a table, transform and load your scraped data into it.

2. Handle Missing Emails & Websites (Medium - but most important)

For jobs with missing emails:

- Check the Application Link for contact details.
- If an external company website is available, attempt to scrape from the homepage or common contact pages (e.g., /contact, /kontakt, /imprint, /impressum) for an email address.

3. Solve CAPTCHA Challenges (Hard)

Explore approaches such as ML/AI-based solutions (train/use OCR model) or third-party CAPTCHA-solving services to automate the process.