Data acquisition, extraction, and storage

Project description

Group members:

- Marwa Nair
- Rezkellah Fatma-zohra
- Kasmi Abderrahmane

Description:

Our project will focus on creating a clean and well-structured dataset of job offers from various websites. We aim to acquire job listings from multiple sources such as job boards, company websites, and social media platforms. The sources may include popular platforms like Indeed, Glassdoor, LinkedIn, and others accessible via APIs or through web scraping.

The key steps of our project will be:

1. **Data Acquisition**: We will use a combination of web scraping and APIs to gather job postings. This step may involve handling challenges such as rate-limiting on APIs and designing robust scraping solutions for websites with dynamic content.
2. **Data Transformation and Restructuring**: The acquired data will likely be in various formats and structures. We will restructure and integrate the data to form a uniform dataset. Tasks may include normalizing different formats (e.g., JSON, HTML), extracting relevant fields (e.g., job title, location, company, salary, requirements), and combining multiple datasets.
3. **Data Storage**: We will evaluate and select an appropriate data storage solution. Our choice will depend on the nature of the data, and we will likely consider options such as relational databases (e.g., MySQL) or NoSQL solutions (e.g., MongoDB) depending on scalability and query requirements.
4. **Data Quality Assessment**: We will ensure the quality of the final dataset by addressing missing or duplicate values, handling outliers, and performing data validation. Both automated tools and manual sampling methods will be used to verify the completeness and accuracy of the data.

The outcome of this project will be a dataset that can be used for various applications, such as building a job search engine, training a machine learning model for job recommendation, or constructing a graphical interface for navigating job offers.