🕵️ WikiDetective: Following the Footprints of Knowledge

CSO Technical University of Munich

This IDP explores how to access and analyze German Wikipedia articles with a focus on structured and time-based content extraction. The project aims to identify the most efficient methods — such as API usage or web scraping — to retrieve key components of an article (e.g., title, introduction text, publication date) as well as its version history over time. A specific focus lies on comparing different extraction strategies in terms of performance, completeness, and sustainability. A key task is to strategically sample from various types of Wikipedia content, such as “Exzellente Artikel,” and to analyze how extraction quality varies across sampling methods and article types. The project will contribute to a broader understanding of how encyclopedic content evolves and how it can be programmatically accessed for research purposes.

🎯 Goals

- **Implement an efficient and reproducible workflow for extracting structured content from German Wikipedia articles** (e.g., title, full article text, publication date, version history, article history)
- **Implement and test multiple extraction methods**, such as MediaWiki API, Wikipedia Dumps, Web scraping, etc..
- **Compare and evaluate these methods** with respect to Data completeness and accuracy, ease of acess and automation and Long-term maintainability and scalability
- **Develop effective sampling strategies** to select representative sets of articles (e.g., "Exzellente Artikel" vs. average articles)
- **Provide structured, research-ready output datasets** (e.g., JSON or CSV format) or **direct export to sql database**
- **Ensure reproducibility and extensibility** of the workflow for integration into future research projects (e.g., media bias analysis)

🎓 Profile

- Java and Python programming experience
- Working with RESTful APIs (e.g., MediaWiki API)
- Handling JSON/XML data
- Basic data processing experience
- **Bonus:** Java/JavaEE & SQL or lightweight database usage (e.g., SQLite)

📄 Deliverables

- Fully functional and well-documented source code for data extraction
- Structured output files (e.g., JSON, CSV) that include title, intro, full text, publication date, version info, and category
- Code comments and a short summary document describing the methodology and comparison of approaches

💪 Who We Are

yathos is a software development and consulting company specializing in **custom software** for research and businesses. We deliver **reliable, low-maintenance solutions** – from consulting to implementation and operation. The **Chair for Strategy and Organization** conducts **impact-driven research** on **future-shaping topics** such as **Agile Organizations, Digital Transformation, Blockchain, Innovation, HRTech, and EdTech**. We focus on **emerging trends** that define tomorrow’s strategies, technologies, and organizations.

How to apply?

Please send an e-mail to ***joe.yu@tum.de***. The e-mail should include (1) the CV and transcript.