RAG-Data & Test

Infomatik

PROJEKTIDEE

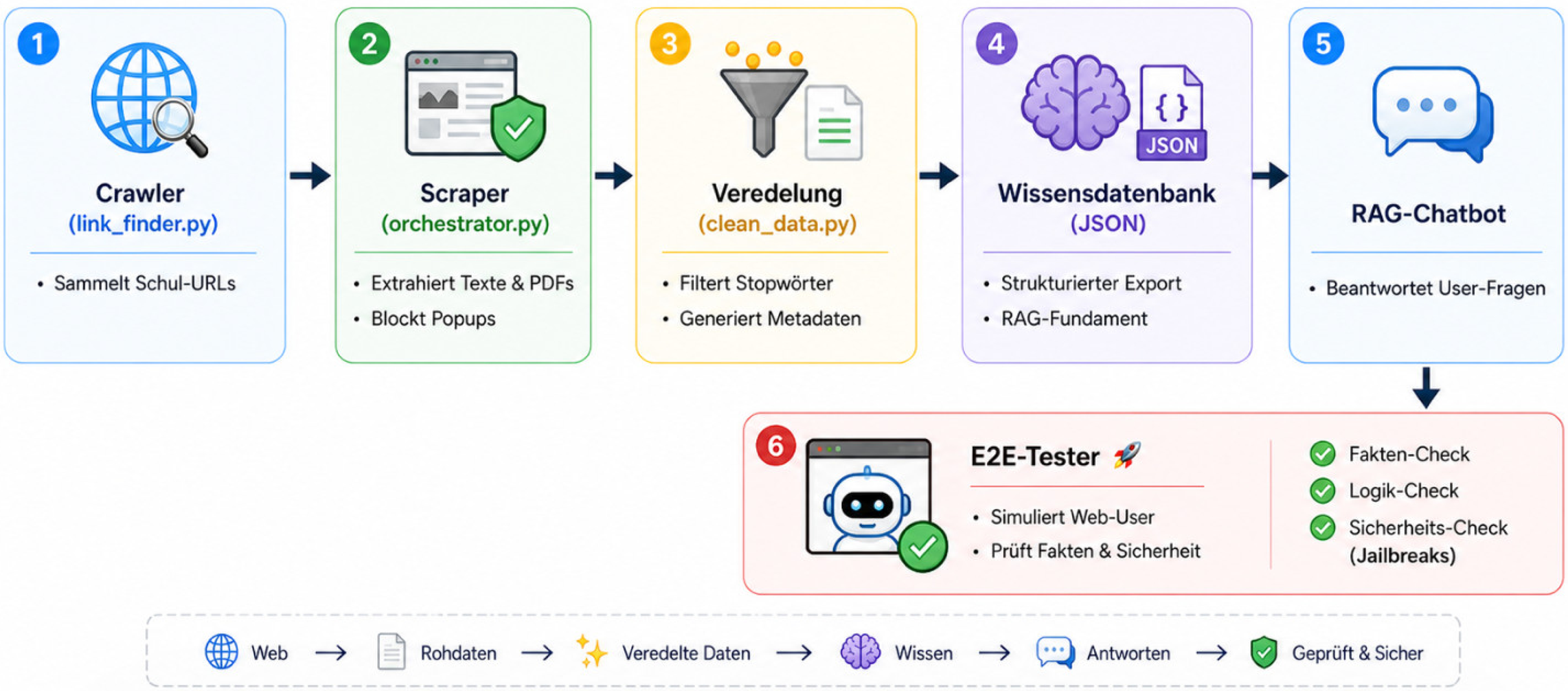
Dieses Projekt bildet das Daten-Backend und die Qualitätssicherung für den KI-Chatbot der TFO Bozen. Ziel ist eine robuste Pipeline, die automatisiert Informationen und PDFs von der Schulwebsite extrahiert und für ein RAG-System (Retrieval-Augmented Generation) aufbereitet. Zusätzlich wurde ein automatisierter Selenium-Testbot entwickelt, der den finalen Chatbot kontinuierlich auf Korrektheit und Sicherheit (Jailbreaks) prüft. Dies liefert das Wissensfundament für die gesamte KI-Anwendung.

FUNKTIONSWEISE

- Crawler durchsucht die Schulwebsite und sammelt URLs
- Scraper extrahiert saubere Texte und PDFs (umgeht Popups/Banner)
- Skripte veredeln die Rohdaten (Keywords, Zusammenfassungen) für RAG
- Dynamischer Export der Daten in eine strukturierte JSON-Wissensdatenbank
- Selenium-Testbot simuliert reale Chat-Eingaben im Browser
- E2E-Testing validiert die Chatbot-Antworten auf Fakten und Sicherheit

DATEN UND FAKTEN

- Kerntechnologien: Python, BeautifulSoup, PyPDF2
- Scraping: Headless-Selenium mit integrierter DOM-Bereinigung ("Popup-Killer")
- Datenaufbereitung: Automatisierte Stopwort-Filterung und Metadaten-Generierung
- Testing-Framework: Dynamische Wartezyklen für asynchrone Web-Frontends
- Qualitätssicherung: E2E-Tests decken Fakten, Logik und Prompt-Injections (Jailbreaks) ab



Projektteam:
Alex Pizzedaz

max valier
TFO BOZEN