

Webscraping with Selenium and BeautifulSoup

Thomas Jonas

2019-06-16

Ziel

- ▶ Automatisiertes Webscraping
- ▶ Datenhaltung in Zeitreihen-Datenbank
- ▶ Darstellung durch Dashboard
- ▶ Deployment in Docker-Umgebung

Inhaltsangabe

- ▶ Übersicht genutzter Technologien
- ▶ Implementierung & Konfiguration
- ▶ Probleme
- ▶ Fazit

Webscraping

Selenium

- ▶ Webscraping von dynamischen single-page Webseiten.
- ▶ Nutzt Browser (chrome-driver muss installiert sein.)

BeautifulSoup

- ▶ Suche nach Klassen und Tags innerhalb HTML-Code.
- ▶ Light-weight

Zeitreihen Datenbank

influxDB

- ▶ Schreiben/Lesen von Zeitreihen-Daten.
- ▶ Meist genutzte ZR-DB.
- ▶ Push-basiert.

Prometheus

- ▶ Sammelt Metriken über zB CPU-Auslastung
- ▶ Pull-basierte Datenbank

Dashboard

Grafana

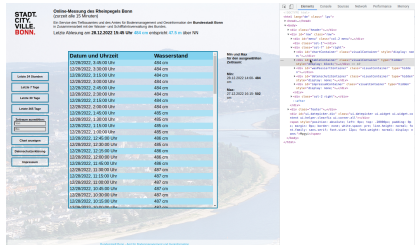
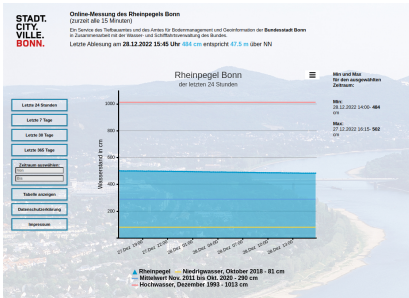
- ▶ Interactive Web-application.
- ▶ Erstellen anschaulicher Graphen und Diagramme.
- ▶ Mehrere Datenquellen möglich.

Deployment

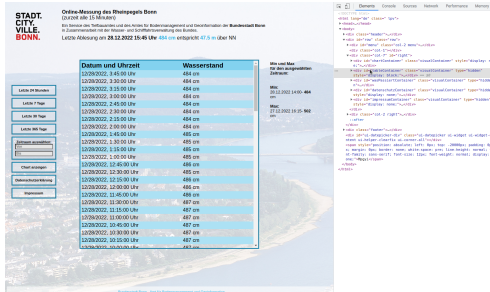
Docker

- ▶ OS-level Virtualisierung. Verpacken von Software in Container.

Implementierung: Webscraping



Implementierung: Webscraping



- Selenium ermöglicht Navigation auf Webseiten.
- Auswahl, Ausführung und Änderung von Elementen.

Genutzte Technologien