Tübix 2019

Webcrawling mit scrapy und PostgreSQL backend

06. Juli 2019

Janek Schoffit

Lightning talk

SCRAPY

Scrapy

- Open source web spider framework
- Erweiterbar durch Plugins und Python libs
- Mit scrapyd als daemon deploybar

Middleware

- Protokolldateien sind lokal auf den Servern gespeichert
- Hoher Aufwand bei manueller Auswertung
- Ereignisse sind nur schwer durchsuchbar
- Auswertungen finden nur im Ernstfall statt

POSTGRESQL

PostgreSQL

- Relationale Datenbank
- JSON Felder mit Indexierung
- Fulltext search support

Suche

- ts_vector

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT