Data Science & Real-Time Big Data Analytics
in cooperation with the start-up Promonode at FZI Forschungszentrum Informatik

Scalable web scraping platform with NLP analytics

Motivation

There are many scenarios where textual information from publicly available websites needs to be analyzed in an automated and scalable way. For many companies (SMEs), it is just not feasible to develop such a web scraping infrastructure from scratch. It would require expertise in various fields such as machine learning, natural language processing, distributed system and also a significant upfront time investment.
As a startup (Promonode), together with you and your team, we would like to explore what it takes to develop such an infrastructure.

Your tasks

As a first step, we will, of course, explain all the necessary parts to your team. This means, you will know exactly what your goals are and how to achieve them efficiently.

You will then be able to develop a distributed stream-processing architecture which can be used for scraping content from various web sites and to analyze the content through an NLP pipeline.

Your work comprises of the following parts:

- Scalable worker queue (e.g. based on Apache Kafka)
- Processing nodes (e.g. based on kafka-streams written in Scala / Javascript / Python or Java)
- Persistence / database layer (scalable NoSQL database)
- Simple platform monitoring view (web or command line based)
- Optionally, dockerization of the components

We offer

- Expertise in scalable stream processing technologies, machine learning and NLP.
- Intensive supervision of the team and valuable feedback.
- Spirit of a growing startup which provides you with a valuable experience.
- Real-world use case and data rather than a toy example.
- Nice working environment with Tischkicker @ FZI

How to apply?

- E-mail: **knoell@fzi.de**