LMU ReMLAV Project
HiWi-Position
CommonCrawl Daten & Elasticsearch

Implementation of a search engine for document retrieval

07.11.2018

Project participant / supervisor:

Prof. Dr. Thomas Seidl	Prof. Dr. Hinrich Schütze	Prof. Dr. Volker Tresp
Michael Fromm	Dietrich Trautmann	Evgeniy Faerman

Intro:
The development of a data set in the area of argumentation mining makes it necessary to be able to search (Elasticsearch) for terms on a data basis (CommonCrawl).

Tasks:

- Realization and configuration of a search engine for document retrieval for any search terms
- The data basis is CommonCrawl (CC, http://commoncrawl.org/the-data/get-started/).
- The search engine is Elasticsearch (ES, https://www.elastic.co/products/elasticsearch)
- Indexing of the entire CC is planned.
- Start of system/pipeline creation on August data: http://commoncrawl.org/2018/08/august-2018-crawl-archive-now-available/
- Reduction to the English text data
- Configuration of the system as a distributed system to scale to the entire CC later
- Need to distribute it across multiple virtual machines (VM) or containers (docker)
- Computing capacity is provided via the LRZ Cloud or AWS
- Weekly communication of the progress, as well as documentation of the code (Gitlab)

Requirements:

- First knowledge with Big-Data technologies
- Programming knowledge in Java and/or Python
- Independent operation under guidance
- Successfully completed B.Sc. in Computer Science and/or related degree programs

Period:

- Start at the next possible date
- Duration: 3-6 months
- Working time: 8h - 20h per week

Salary according to official LMU auxiliary staff rates

Please send your application with proof of performance and other relevant documents:
dietrich@cis.lmu.de or fromm@dbs.ifi.lmu.de