Corpus Processing Service: A Knowledge Graph Platform to perform deep data exploration on corpora.

Peter Staar[1], Michele Dolfi[1], and Christooh Auer[1]

[1]IBM Zurich Research Laboratory

September 16, 2020

Abstract

Knowledge Graphs have been fast emerging as the {\em de facto} standard to model and explore knowledge in weakly structured data. Large corpora of documents constitute a source of weakly structured data of particular interest for both the academic as well as the industrial world. Key examples include scientific publications, technical reports, manuals, patents, regulations, etc. Such corpora embed many facts that are elementary to critical decision making or enabling new discoveries. In this paper, we present a scalable cloud platform to create and serve Knowledge Graphs, which we named \textit{Corpus Processing Service}. Its purpose is to process large document corpora, extract the content and embedded facts, and ultimately represent these in a consistent knowledge graph that can be intuitively queried. To accomplish this, we use state-of-the-art natural language understanding models to extract entities and relationships from documents converted with our previously presented CCS platform. This pipeline is complemented with a newly developed graph engine which ensures extremely performant graph queries and provides powerful graph analytics capabilities. Both components are tightly integrated and can be easily consumed through REST APIs. Additionally, we provide user-interfaces to control the data ingestion flow and formulate queries using a visual programming approach. The CPS platform is designed as a modular microservice system operating on Kubernetes clusters. Finally, we validate the quality of queries on our truly end-to-end knowledge pipeline in a real-world application in the oil and gas industry. To date, the capabilities of CPS are successfully leveraged in more than 5 client engagements.

Hosted file

`wileyNJD-AMA.pdf` available at https://authorea.com/users/359305/articles/481396-corpus-processing-service-a-knowledge-graph-platform-to-perform-deep-data-exploration-on-corpora