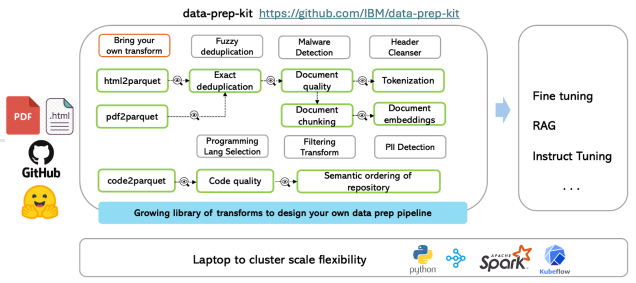
Preparing Data for LLM Applications Using Data Prep Kit

.ical

12-04, 17:30–19:00 (UTC), LLM Track

Data Prep Kit is a new open source python library to help you wrangle and clean your data for generative AI applications (de-dupe, detect language, removing PII, detect malware, creating embeddings, etc.)



When building data intensive applications, a significant portion of your time will be dedicated to data wrangling (cleaning, de-duping, removing markups, etc.). Data Prep Kit (https://github.com/IBM/data-prep-kit) is a new open source python library that can scale from your laptop to a highly scalable cluster in the cloud. It has been used at scale to prepare terabytes of data to train the IBM Granite Large Language Models (LLMS). A few noteworthy features of DPK include: de-duping documents (exact dedupe and fuzzy dedupe), handling documents and code, language detection (spoken languages and programming languages), removing PII, malware detection and creating embeddings for a vector database. In this talk, Sujee will go over some interesting features of the Data Prep Kit and show a demo.

Prior Knowledge Expected –
No previous knowledge expected

Dave Nielsen

Results-oriented Developer Relations professional with 20 years in open source, databases, devops, payments and AI, and now 1 year in generative AI. Proven ability to build and manage vibrant, engaging developer communities, drive developer adoption of cloud products, and foster strategic partnerships within the developer ecosystem. Skilled in community building, presenting technical topics, communication and content creation. Passionate about developer experience and success.

Sujee Maniyam

Sujee Maniyam is an expert in Generative AI, Machine Learning, Deep Learning, Big Data, Distributed Systems, and Cloud technologies. He is passionate about developer education, fostering community engagement. Sujee has led numerous training sessions, hackathons, and workshops. He is also an author, open source contributor and frequent speaker at conferences and meetups.

portfolio • Linkedin • Github