Richtlinie zum „Software-Sprint“

ML Tooling

Schlussbericht

Zuwendungsempfänger:

Kalkan Masuch Raethlein GbR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01|S20S49 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Die Digitalisierung stellt durch Kosten und fehlendes Know-how, besonders in komplexen Gebieten wie Infrastruktur und Data Science, vor allem kleine und mittelständische Unternehmen vor große Herausforderung, wodurch die Innovationskraft gehemmt sein könnte.

Durch hohe Komplexität beziehungsweise die Abhängigkeit von bestimmten Plattformen besteht also die Gefahr, dass nur Unternehmen mit entsprechenden Ressourcen von Machine Learning und somit von intelligenten Systemen profitieren.

Die sinnvolle Kombination und Erweiterung von bestehenden Open-source Tools kann dabei helfen diese Technologien einer breiteren Zielgruppe zugänglich zu machen. Daher umfasst das Projekt ein integriertes Toolset, um den Einsatz von Machine Learning zu erleichtern und das mächtige, aber komplexe, Ökosystem von Open-source Tools einfacher zugänglich zu machen. Das Toolset besteht aus mehreren kompatiblen Komponenten, welche auf populären Open-source Tools aufbauen und durch sinnvolle Erweiterungen ergänzt werden. Diese Komponenten ermöglichen die effiziente Nutzung von vorhandenen Server-Ressourcen, die einfache Verwendung eines modernen Data Science Technologiestacks, das Synchronisieren von Experimenten und das einfache Deployment von Machine Learning Modellen für den produktiven Einsatz. Der Data Scientist kann sich daher auf seine Aufgabe fokussieren, anstatt viel Zeit für notwendige Infrastruktur Aufgaben investieren zu müssen.

Die wichtigsten Meilensteine waren, Dokumentationen und automatisierte Tests für vorhandene Komponenten bereitzustellen um die Nutzung und Wartbarkeit zu erhöhen, vorhandene Komponenten zu updaten, zwei neue Komponenten zu entwickeln, eine für das Tracken von Experimenten und eine für das Starten von Machine Learning Services und dann über das Verbreiten von Blogsposts und ähnlichen Inhalten über Onlinemedien Aufmerksamkeit für das Projekt zu generieren und die Community weiter wachsen zu lassen.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Moderne Techstacks können recht komplex sein. Zum Beispiel sollten Data Scientists sich im besten Fall komplett auf ihre Aufgaben, das Experimentieren mit Daten und Entwickeln von Lösungen auf den gesammelten Erkenntnissen, konzentrieren können. In der Realität sind Teams allerdings klein und einzelne Personen müssen eine Vielzahl von Technologien und Konzepte beherrschen. So müssen Data Scientists nicht nur in der Lage sein, zum Beispiel die Programmiersprache Python zu beherrschen, sondern sich auch mit Fragen der Skalierung, der Datenbeschaffung und -verteilung, dem Aufsetzen der nötigen Infrastruktur und dergleichen befassen. Mit unseren Komponenten versuchen wir, Komplexitäten zu abstrahieren, damit sich die Nutzenden auf ihre eigentlichen Aufgaben und Themengebiete konzentrieren können.

Ausführliche Darstellung der Ergebnisse

- Erreichte Meilensteine:
    - Wir haben für einige unserer Komponenten automatisierte Workflows via GitHub Actions eingerichtet und darüber automatisierte Tests ermöglicht. Dies hilft bei der nachhaltigen Wartbar- und Nutzbarkeit der Projekte durch uns und die Community. Die Erkenntnisse und unsere Best-practices haben wir ebenfalls in eine dedizierte Software Bibliothek (universal-build) integriert, dokumentiert und ebenfalls veröffentlicht.
    - Unsere Hauptkomponente, ML Workspace, haben wir geupdated, das schließt unter anderem beinhaltete Tools und ML Bibliotheken ein. Das ML Workspace Projekt hat im Förderzeitraum ca. 700 Sterne auf GitHub dazubekommen und wurde insgesamt über 100.000-mal von DockerHub heruntergeladen (fein-granularer gibt es die Daten leider nicht).
    - Wir haben mit der Entwicklung einer Experimententracking-Komponente begonnen. Die Webapp und Serverkomponente sind in einer ersten Version programmiert worden.
    - Wir haben ein best-of Projekt gestartet, worüber wir viel Aufmerksamkeit für die GitHub Projekte bekommen haben und welche uns selbst dabei hilft, die richtigen Tools und Bibliotheken zu finden, die in zukünftigen Versionen des ML Workspaces installiert sein sollten.
    - Die Python Bibliothek für das Experimententracking sowie die Unified Model Bibliothek zum Einfachen Starten und Verwalten von Machine Learning Models konnten leider während des Förderzeitraumes nicht fertiggestellt werden, werden aber noch veröffentlicht.
- Wesentliche Erfahrungen:
    - Wir haben durch das Projekt den Schritt in die Selbstständigkeit gewagt und evaluieren nun, nach der Förderung ein richtiges Startup aus der Idee zu gründen.
    - Wir haben neben der Arbeit an den Komponenten mit Fokus auf Pythonprogrammierung viel über CI/CD Pipelines und Testing gelernt.

- Durch Fokus auf Pythonprogrammierung haben wir uns Verständnis der Programmiersprache Python stark vertieft, z.B. dessen Concurrency/Parallel Funktionalität, event-basierte Asynchronität etc.
- Durch die Coronasituation haben wir unsere Arbeitsprozesse und -tools auf Remote und Async-first optimiert.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Die Tools zielen auf Data Scientists und Machine Learning Engineers ab, welche alleine oder in Teams arbeiten. Das Toolset zielt dabei vor allem auf einzelne Entwickler und kleine Entwicklerteams, Start-Ups, Klein-und Mittelstandsunternehmen, Universitäten, öffentliche Einrichtungen und NGOs ab.

Dank der Containerisierung der Komponenten und Abstrahierung der unterliegenden Komplexitäten (z.B. Einrichten von SSH Verbindungen usw.) soll die Zielgruppe in die Lage versetzt werden, schnell und einfach eine homogene Machine Learning Entwicklungslandschaft aufzusetzen. Somit benötigen die Nutzenden weniger Fachwissen in Bereichen wie zum Beispiel System Administration im Bereich SSH oder Betriebssysteme ohne dabei auf proprietäre Services setzen zu müssen und sich somit stark von anderen Dienstleistern (z.B. Cloudservice Anbieter) abhängig zu machen. Die meisten Komponenten haben ein sogenanntes *Quickstart* Tutorial, damit sie schnell und einfach eingesetzt und ausprobiert werden können. Während der Förderung gab es eine aktive Mitwirkung der Community in Form von Feedback, Issues und Pull Requests auf GitHub, was den Nutzen von Open-source Technologie unterstreicht. Darüber hinaus erhielt das best-of Projekt viel Aufmerksamkeit (in Form von Github Sternen) von der Community.

Wir planen das Projekt fortzuführen, in dem die Komponenten beworben werden und so andere Teams, Firmen, und Projekte dazu bringen, sie einzusetzen. Des Weiteren evaluieren wir, ein Startup zu gründen und die Komponenten zu einer Plattform zusammenzuschließen, welche als Produkt angeboten werden könnte.

Wie schon in der vorherigen Sektion erwähnt, haben wir uns durch das Projekt in den Bereichen Selbstständigkeit und Remotearbeit, sowie in Python, CI/CD und Testing stark weiterentwickelt.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Zwei Komponenten, die wir während des Förderzeitraums entwickeln wollten, werden erst nach der Förderungszeit fertiggestellt. Aus langfristiger Sicht hat es Sinn gemacht, einige Tätigkeiten vorzuziehen. Für die Experimententrackingkomponente ist die Webapp und der Server fertig, allerdings fehlt noch der Python Client. Die Serviceverwaltungskomponente um Machine Learning Models einfach zu deployen konnten wir während des Förderzeitraumes leider nur konzeptionell starten. Die anderen Tätigkeiten wie zum Beispiel das Einrichten von automatisierten Testpipelines basierend auf Containern hat mehr Zeit als ursprünglich geplant eingenommen. Uns war es allerdings wichtig, diesen Tätigkeiten den erforderlichen Fokus zu widmen um die Projekte auch nachhaltig wart- und nutzbar zu machen.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Die Ergebnisse können in den entsprechenden Repositories der GitHub-Organisation eingesehen werden: https://github.com/ml-tooling

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Die Qualitätsarbeit im Hintergrund wie zum Beispiel CI/CD Pipelines, Tests, Codequalitätstools etc. hat anfänglich mehr Zeit in Anspruch genommen als gedacht. Das hat zur Verzögerung bei der Entwicklung einiger Komponenten geführt. Die Ergebnisse dieser Tätigkeiten wurden ebenfalls veröffentlicht und sauber dokumentiert, so dass auch potentiell jeder andere davon profitieren kann.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

- Wir haben während des Zeitraums einigen Input von der Community für unser Projekt bekommen und wurden im Rahmen unserer Verbreitung der best-of Listen, um Aufmerksamkeit für die Projekte zu gewinnen, von einem Podcast erwähnt und haben auf Twitter und Reddit eine große Zielgruppe angesprochen, sodass wir auf diesen best-of Listen insgesamt während des Förderzeitraums ca. 5000 Sterne auf GitHub bekommen haben.
- Zusammenarbeit mit einem Berliner Startup und einem Team eines großen Unternehmens um Ideen auszutauschen und zu validieren. Durch diese Gespräche haben wir unsere Roadmap zwischenzeitlich etwas angepasst.
- Wir benutzen für unsere Komponenten und in unserem Entwicklungsprozess selbst einige Open-source Tools, welche uns sehr weitergeholfen haben. Während des Förderzeitraums haben wir zu ein paar Komponenten auch in Form von GitHub Issues oder Pull Requests contributed.