ENTERPRISE DATA MANAGEMENT

TRAINER - ACTIVE LEARNING

3. SPRINT - 09.07.2019

TOBIAS KLATT, JACQUELINE HOFFMANN, MARIE SCHUMACHER UND TABEA FRENZEL

RECAP ❓

- Named Entity Recognition und Machine Learning Aufgaben erfordern eine Vielzahl annotierter Trainingsdaten
- manuelles Annotieren von Daten ist aufwendig und kostenintensiv

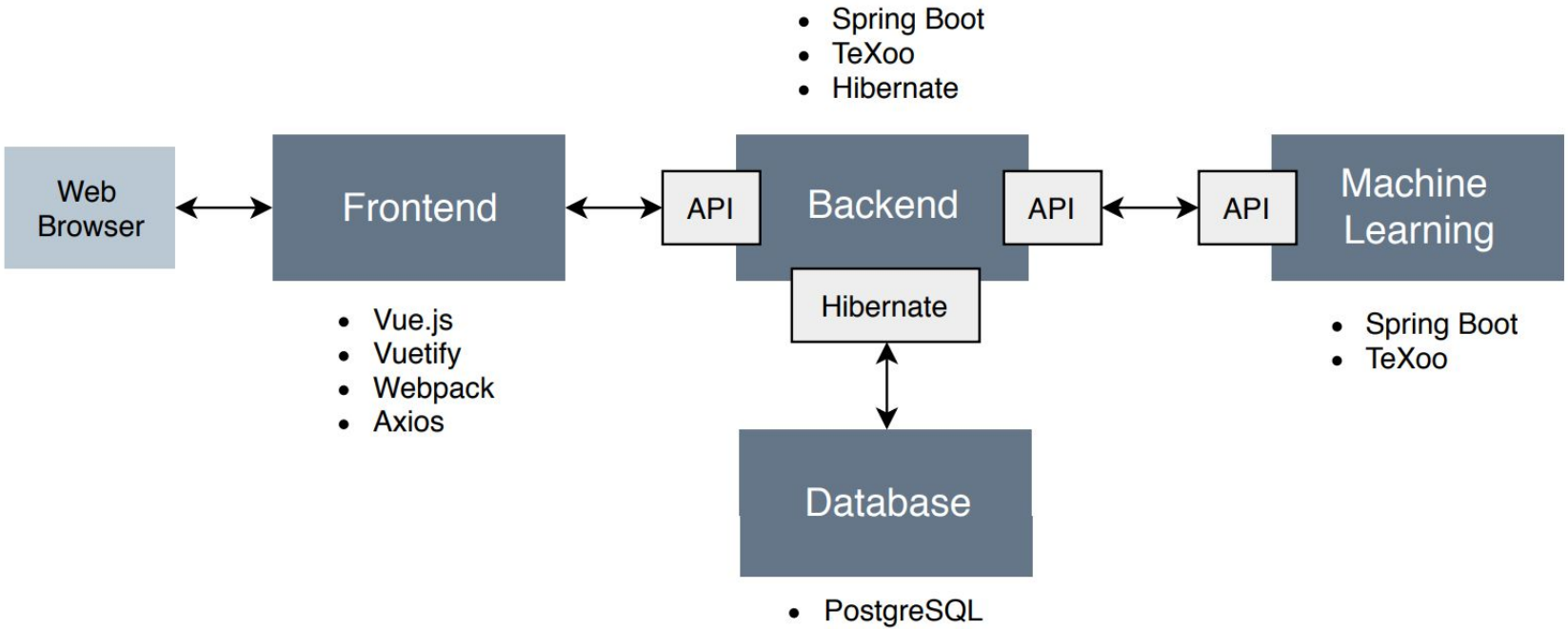
➔ TRAINER

TraiNER 💻

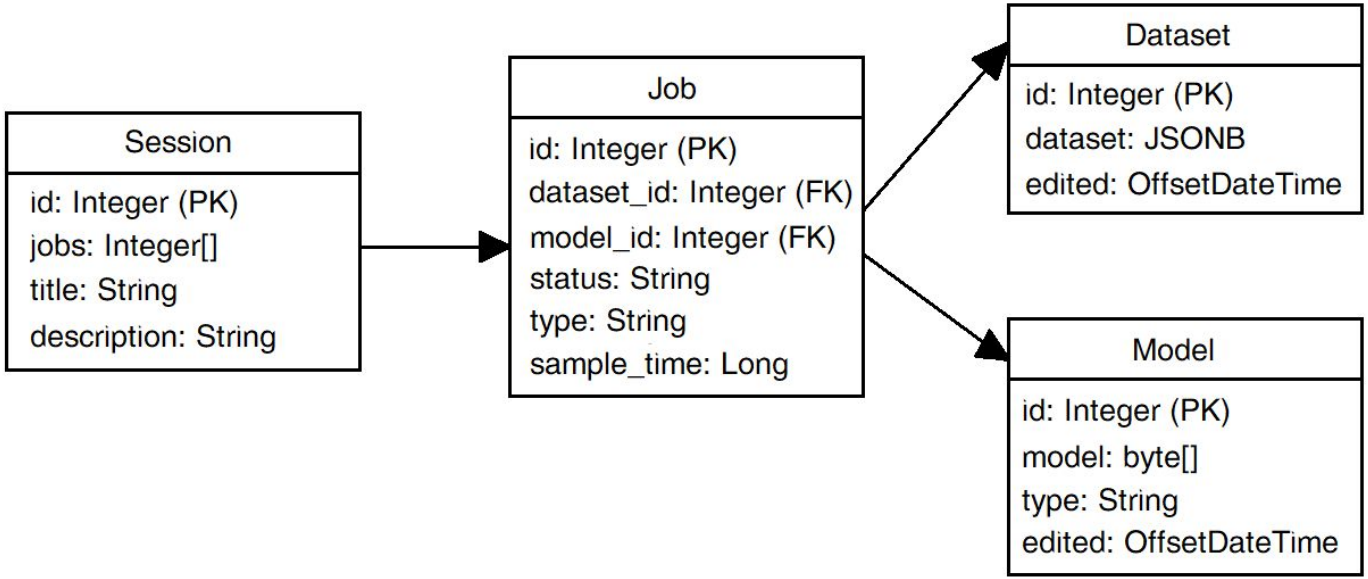
- mit austauschbaren Samplingstrategien soll Maschine selbst entscheiden, welche Daten durch Menschen annotiert werden müssen
- Modell ohne großen manuellen Trainingsaufwand verbessern

→ ACTIVE LEARNING

Architektur 🏠



Datenmodell 💾



Ablaufdiagramm🔄

Link zum Ablaufdiagramm

Sampling

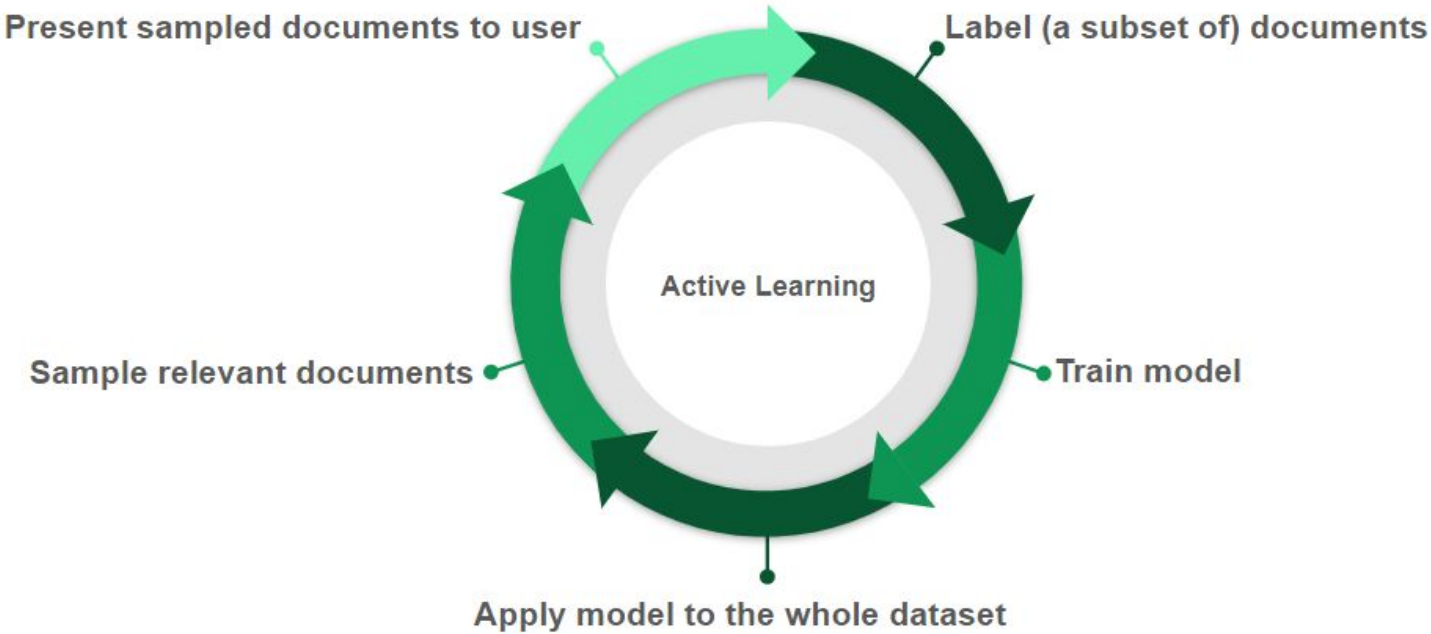
Sampling Strategien

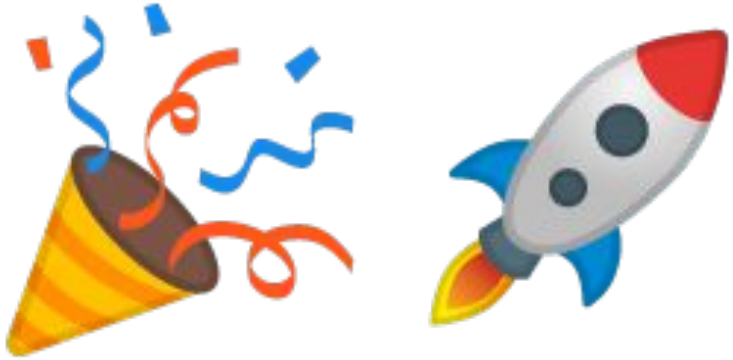
- Random Sampling
  - zufällig ausgewählte Samples
- Uncertainty Sampling
  - Samples, bei denen sich die ML (das Modell) am unsichersten ist
  - Nutzer annotiert diese Samples und schickt sie zurück an ML

Upsampling

- Nutzer annotiert und startet das Training
- Annotationen werden auf restlichen Datensatz angewendet

Ziele für diesen Sprint 🏁





Umgesetzte Aufgaben 📚

Aufgabe	**Zuständigkeiten**
API Anpassungen	gesamtes Team
Beschreibung für Sessions	gesamtes Team
Metriken zum Training	gesamtes Team
Sampling Strategien	Marie, Tabea
Pagination für Samples	Tobias
Samples hochladen	Marie, Jacqueline
Upsampling	Marie

Umgesetzte Aufgaben 

Aufgabe	**Zuständigkeiten**
Starten eines Trainings	Tabea, Jacqueline
Model herunterladen	Marie, Jacqueline
Nutzerfeedback	Jacqueline, Tobias
Route Guards	Jacqueline, Tobias
Userguide und Dokumentation	gesamtes Team
Dockern der Anwendungen	Tabea, Jacqueline

Probleme ❗

- wenig Erfahrungen in der **Backend-Entwicklung**
- Integration **TeXoo**
- viele **API-Anpassungen**
- **Kommunikation** von
  - der Frontend- mit der Backend-Komponente
  - der Backend- mit der Machine Learning-Komponente
- **Docker** Container aufsetzen

Ergebnisse 🌟

- **Weiterentwicklung** Projekt TraiNER
- Backend mit austauschbaren **Samplingstrategien**
- neue **Features** im Frontend
- Bereitstellung im **Docker**

→ **Trainingsdurchlauf** konzipiert & implementiert

Lessons Learned 📚

- **Probleme** während Implementierung eines **Trainingsprozesses** kennengelernt
- Full-Stack-Entwicklung
- Kennenlernen neuer **Technologien**
- **Sampling-Strategien**

Ausblick - Implementierung 👀

- **Samples**
  - weitere / andere Sampling Strategien implementieren
  - Random Samples, wenn keine Annotationen von ML kommen
- **Nutzerfeedback** verfeinern
- **Tests** schreiben

Ausblick - Messbarkeit 👀 📉

- **Ablauf**
  - Model erstellen mit Datensatz ohne Annotationen
  - Model erstellen mit Goldstandard
- **Messung und Vergleich**
  - Berechnung **F1-Wert** $F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$

Vielen Dank 🙏

Quellen 📚

- https://www.pexels.com/photo/coffee-writing-computer-blogging-34676/ (Stand: 24.04.19)
- https://docs.google.com/presentation/d/10uGrGzN4cdL_zpyn-OTvdu3KsNfjUpb_4mkXiBI6pkw/edit#slide=id.g55f59cb9b8_0_268 (Stand: 08.07.19)

Zusatz - Vergleich

TraiNER am Anfang des Semesters	TraiNER heute
• Backend-Stub, kein Active Learning • Funktionen: Annotieren • Trainingsdurchlauf mit Active Learning nicht möglich	• Funktionierendes Backend, Active learning • Funktionen: Sessions, Metriken, Up- und Download, Training starten • Trainingsdurchlauf mit Active Learning möglich