Maximilian Kähler, Nico Wagner | DNB, Leipzig

Datenmanagement mit DVC

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

Das Datenmanagement und Pipelining-Tool DVC[1]

Kernfunktionen:

- **Versionsverwaltung** allá git für (große) Dateien
- Unterstützung von Machine-Learning Workflows durch **reproduzierbaren Pipelines**
- **Teilen und Synchronisieren** von Pipelines und Daten im Team

[1] dvc.org

Use Case: Data Registry

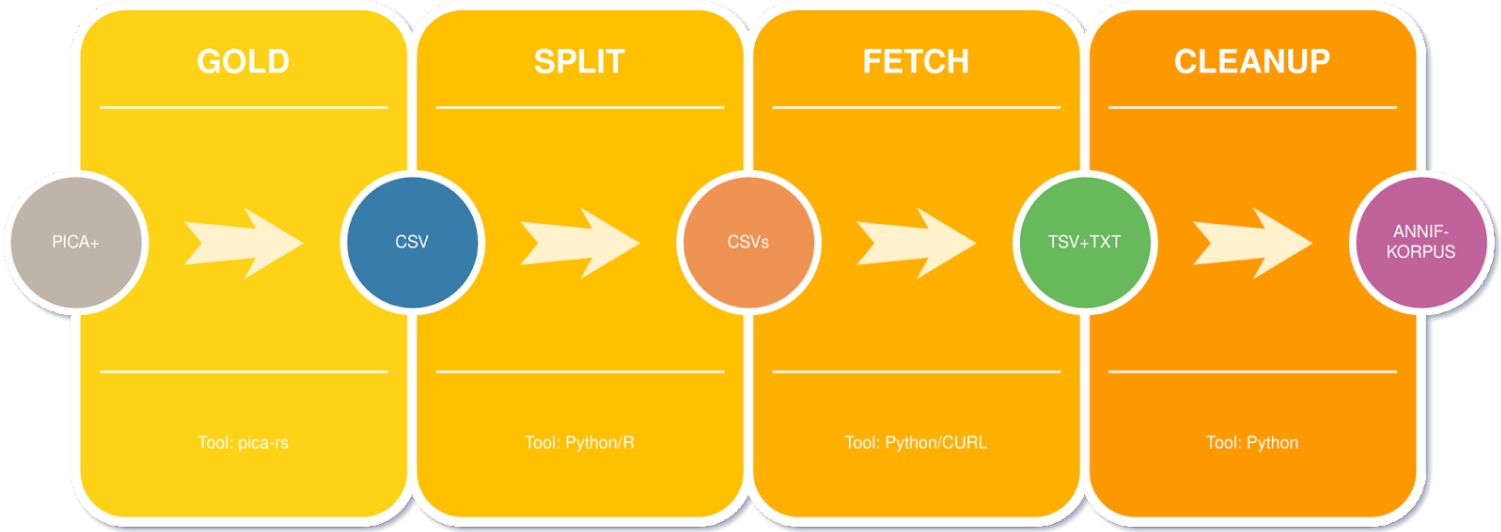
Grafik aus Urheberechtsgründen entfernt. Siehe Link

Quelle: https://dvc.org/doc/use-cases/data-registry

Der DVC-Storage – „das Archiv“

- Storage dient zur permanenten Ablage von Daten
- Anbindung mehrerer Storages möglich
- Storage-Typen: Verzeichnis, SSH, S3, Google Drive, …
- Solide, angemessene Infrastruktur nötig (backups, schnelle Netzwerkanbindung, große Festplatten)

Use Case: Pipelining Tool



Use Case: Pipelining Tool

- DVC garantiert Reproduzierbarkeit von Pipeline
- DVC spart überflüssige Wiederholungen von Stages
- Metriken, Parameter & Plots überwachen den Input und Output von Pipeline
- Verwaltung von Experimenten

Der DVC-Cache – „der Zwischenspeicher“

- Speichert Artefakte auf einem Server zwischen
- Verhindert unnötiges Wiederausführen von Pipelines
- Geteilte Caches vermeiden Redundanz
- Muss regelmäßig bereinigt werden

Einsatz in der DNB

- Verwaltung aller Daten und Schritte, die für die Modellerzeugung nötig sind
- Bereitstellen und Teilen von Modellen und Textkorpora
- Aufbau von Pre- und Postprocessing-Pipelines
- Archivierung von Stammdaten (PICA-Abzug, CSV-Dateien)

Erkenntnisse (1)

- Kollaboratives Arbeiten an Datasets ohne Sorge etwas „kaputt“ zu machen
- Erleichtert Datenaustausch mit anderen Mitarbeiter*innen

Erkenntnisse (2)

- Pipelines werden sehr schnell komplex und unübersichtlich
- Annif-Korpusformat ungünstig (inodes-Problem)
- Performance-Probleme bei vielen Einzeldateien (MD5-Problem)

Nächste Schritte

- Konsolidierung und Stabilisierung der Python-Tools für Datenmanagement
- Umstellung auf neue Korpusformate
- Implementierung der kompletten Aufbereitungs- und Trainingspipeline in DVC
- Hyperparameteroptimierung mit DVC
- Portierung der Pipelines auf HPC

Vielen Dank!

Fragen und Anregungen jederzeit an:

Maximilian Kähler, Nico Wagner

E-Mail: m.kaehler@dnb.de, n.wagner@dnb.de

Unser DNB-KI-Projekt: https://www.dnb.de/ki-projekt