Anwesenheit: Paul, Naomi, Robin, Philipp, Timon

Fancy-ml.dodis.ch

Unser schritt: spacy

Gewisse stellen im Text markiert, dann diese Stellen mit Datenbank abgleichen.

Welcher Müller? (ML)

➔ Output: TEI-XML

Hauptproblem: keine Eindeutigkeit beim Linking.

Ziel: NEL Komponente austauschbar machen, nicht nur für Dodis

- Wikidata NEL Komponente
- Custom TEI NEL trainer (TEI als input, Script macht eigener Linker)

Auf webserver keine App

(NEL = linker, NER = recognition)

Tutorial für Historiker*innen die kein Plan von Technik haben

Wahrscheinlich mehrere komponenten (eine für Dodis, eine für Wikidata, eine Dritte)

Out of scope: kein Front end

API's und Tools:

Spacy: ML framework für NLP

"transformer models sind die geeignetsten" ⇔ wod to mec

OCR optical character recognition ist der input aber ist nicht perfekt also ist nicht perfekt.

TEI-XML nicht ein typisches Dateiformat.

Trainieren mit GPU!

oder

Auf „ubelix cluster" trainieren

Erste Interation

1. „NEL Wikidata tutorial lesen"

Wikidata ist riesig, wir sollen ein Filter entwickeln der nur den Teil vom Wikidata erunterlädt, welche wir brauchen

2. Daten loaden (mit diesem Filter)
3. Training vorbereiten

   Wir müssen noch nichts „liefern“ aber zeigen das wir etwas gemacht haben

Wikidata sehr anders zu Dodis, Projekt von Tobias nicht wircklich brauchbar für uns

Einteilung im Team Vorschlag:

- Tutorial lesen alle
- Bei jemandem Laufen lassen

- Data Handling
- Training

Wir brauchen nicht wirklich Test (weil ML)

Gutes ReadMe

Kommentare und Dokumentation einfach nach best practice

Parameter über config files

Assertions über alles

4 Iterationen:

1. Verstehen
2. NEL Wikidata liefern
3. TEI
4. Trainieren

Wenn Wikidata zu gross -> lieber so 10‘000 daten nehmen.

Ganz am anfang mal mit 100, dann ein paar Tausend

Optional in Docker file

Python