Meeting 21.04.2021 Projekt KD

**Scraper Optimierung**

* Namematching Faktor 10 erhöht aber ausgelagert
* TH Seite -> welche Tabs sind relevant?
* News Seite -> Mehr Anzeigen Button wird nicht berücksichtigt
* Projekttab (alle Projekte die gelaufen sind 450 Projekte) Sehr wertvolle Informationen Mitwirkende etc. -> für Recommender Systeme
* Annahme Professor des Projekts leitet kennt sich im Fachgebiet aus
* Aktuell alles gescrapt

**Diskussion**

* Scraper Optimierung -> spezifische Scraper? Oder alles scrapen? Alles = sehr viel Müll
* Alles was einen Namen beinhaltet ist relevant
* Besserer Ansatz ist alles scrapen als spezifisch
* Keine separaten Scraper -> Überprüfung nötig
* Vereinen der Scraper? Es macht keinen Sinn mehrere Scraper zu haben (Mehrere Scripte für ein und dasselbe Ziel)

**Pia**  
Einarbeitung in Fabians Code /Script noch nix gefixt Coworking Pia & Fabian + Matteo

**Sascha**

* Maria DB Zugang läuft Zugangsdaten in Repo hinterlegt (Vorrausetzung aktivierter VPN)
* Gibt’s ein Schema? Können wir in die Datenbank pushen? -> Ja ist hinterlegt

**Fabian**

Lehrgebiete durchgegangen (350Stk) -> auf der Suche nach Erschließungs Tool ähnlich Midos Ziel ca. 700-800 Lehrgebiete

**Jonas**

Json Datei erstellt mit allen Personen der TH (wenn Fakultät zugehörig wird dies den Personen zugeordnet)

Dumb aus Daten der Personen Field of Teaching etc. (schließt an Jonas Topic an)

Erzeugen der Datei via python script der rest händisch Intellektuelle Erschließung

**Leon & Andy**

* Verarbeitungs Pipeline
* Orientierungsleitfaden
* Herausforderung Lizenzen für Germanet

**Saskia**   
Topic Modelling Problematik müssen noch gefiltert werden

**Jüri**

* 2x Pipeline, um den beschrieben Case zu bearbeiten (auf Basis von Ubuntu System)
* Feedback: Script läuft nicht + Erstelltes Dataframe wird nicht angezeigt
* Verwendung von Pyspark -> Anbindung an die Datenbank