Meeting 05.05.2021, 09:45 – 11:15 Uhr

Report

Leon und Andi:

* Dump hochgeladen
* Py-altemy für pull und push; Dataframe in Json, dann SQL umgewandelt
* Tokens aus Titel und Artikel
* Strings definiert und eingebaut
* Sprache wird bestimmt mittels Stoppwörter, dann durch einen weiteren Schritt überprüft
* Stemming mit Porterstemmer von Wordnet
* Lematisierung auch eingesetzt, da Stemming nicht zufriedenstellende Ergebnisse liefert
* Problem: Listen nicht für SQL-DB geeignet, deshalb wurden Strings definiert
* tf/idf Problematik ist rausgeworfen
* Fazit = Pipeline steht also
* Doku muss noch folgen und etwas mehr Struktur (Aufräumen ffW.)

(LDA „ Topics hier nur aus Dok“)

Michelle und Saskia:

* Aktueller Stand: Warten noch auf die Daten (Topics)
* (Info: Jedes Dokument hat mehrere Topic-ID’s)
* To do ffW.:
* Topic-Speicherung (mehrere Topics auf „Sinnhaftigkeit prüfen“ -> daraus soll ein Haupttopic erstellt werden (aus: automatisch erstellten und händisch erfassten Topics s. Fabian G.)
* Skripte für die Erstellung sind vorhanden
* Annäherung über Wort-Vektoren oder Clustering etc. müssen noch durchgeführt werden
* Evaluierung offen

Martin/Jüri UI:

* Basis: HTML+CSS steht so weit
* Erster Entwurf in GitHub für alle einsehbar (<https://github.com/DIS-KD-Project/UI>) Auch im MM-Channel oder auf Miro

Fabian G.:  
- Topics werden händisch erstellt; Noch nicht final

* Topics kommen aus den Lehrgebiet(en) (-> Diskussion wie sinnvoll & Zusammenführen mit LDA-Topics für kommende Woche festgelegt)

Fabian Ax:

Interne Quellen: Namematching angepasst

Personen gefiltert auf relevante Personen = Matching nur noch auf ½ Personenanzahl

Frage klären: Was kann alles gescraped werden?

* Externer Scraper wird verkleinert („paywalls“ werden ausgelassen)
* Repo muss aufgeräumt werden (Task ffW.)

Sascha:

* DB machte bei Sascha keine Probleme
* DB der TH weist leider Probleme auf
* Hr. Brenner wurde kontaktiert
* Die Verknüpfungstabelle lässt sich auf der TH-DB nicht erstellen (Auch Herr Brenner wusste keine Lösung)
* Sascha hat den entspr. Table händisch zur TH-DB hinzugefügt (Export-Import)
* Ein kleines Problem gibt es mit der TH-DB und dem Datenimport aktuell noch = wird geklärt

Erste Ideen für neue Tasks der folgenden Woche:

* Pipeline-Team: Doku nachreichen, Aufräumen,dann Datenbankimport
* Andi und Leon kommen zu Sascha hinzu und unterstützen; Schnittstelle damit nicht Daten in die DB manuell eingelesen werden müssen
* LDA
* Fabian Gitzler: Manuelle Topics
* Scraper: Pia externe Fokus, Dokstrings, Doku und Aufräumen im Repo

Genaueres wird in Miro festgehalten.

Es wurde noch mal um die Erstellung bzw. Vervollständigung der Dokumentation innerhalb aller Gruppen gebeten.