## Work-Breakdown-Structure

### Projektinitiierung

* Projektziele und Scope festlegen
* Erfahrungen zum Thema NLP und Python im Team erfassen
* Tools festlegen
  + Kollaborationsplattform: GitHub
  + Kommunikationsplattform: Zoom, Discord
  + Entwicklungsumgebung: Anaconda
  + Programmiersprache und Bibliotheken analysieren und Dokumentationen sichten: Python, Spacy, OpenVC, Pytesseract, scholarly
  + Versionsverwaltung: GitHub
* Arbeitsweise festlegen: Scrum, WEKAN
* Termine planen

### Anforderungsanalyse

* funktionale Anforderungen spezifizieren
  + Literaturverzeichnis in BibTex umwandeln
  + Webservice
  + Weboberfläche
  + Einzulesende Datenformate: Bilder, Text, PDF
  + Ausschreiben von Akronymen
  + Ausschreiben von Abkürzungen
  + Vervollständigung fehlender Einträge mittels scholarly
* nichtfunktionale Anforderungen spezifizieren
  + Webserver
  + Betriebssystem
  + Datenbanken
  + Datencharakteristika
  + Performance: Mindestwerte Precession, Recall und F1-Score festlegen
* Technische Architektur
* Prozessbeschreibung
* Anforderungen priorisieren
* Glossar erstellen
* Pflichtenheft erstellen

### Datensammlung

* Literaturverzeichnisformate (MLA, APA…) analysieren
* Datenquellen bestimmen
* Methodik: Web Scraping
* Repräsentativer Stichprobenumfang ermitteln
* Datenarchitektur aufbauen
  + Datenformat
  + Ort der Speicherung
  + Speichertechnologien
  + Datenzugriff / Datenschnittstelle

### Explorative Datenanalyse / Datenqualität bestimmen

* Datencharakteristika bestimmen
* Datenattribute bestimmen
* Metrik zur Datenqualität bestimmen
  + Wann „fit for pupose“?
* Fehlertypen bestimmen
  + Duplikate
  + Fehlende Werte
  + Ausreißer
  + Rechtschreibfehler
  + Synonyme
  + Homonyme
  + Integritätsbedingungen / Grammatikfehler
* Erkenntnisse Visualisieren
  + Fehlerverteilung
  + Verteilung Attribute
  + Bias

### Datenbereinigung

* Art von Datenbereinigungen (Imputation…) für Attribute analysieren
* Daten bereinigen

### Datenvorbereitung

* Satzsegmentierung
* Tokenisierung
* Normalisierung
  + Entfernung von Stoppwörtern
  + Stammbildung und Lemmatisierung
  + Entfernung von bestimmten Interpunktionszeichen
  + Kleinschreibung
* Feature Engineering / Text Representation
  + Merkmalsextraktion
  + Merkmalsauswahl
  + Merkmalskonstruktion

### NLP-Modellierung

* NLP-Modelle / NLP-Algorithmen auswählen
  + Ziel des Modells: Named-entity recognition
* Modelle trainieren / validieren
  + K-fold Cross-Validation
* Modelle evaluieren
  + Evaluations-Metriken: Precession, Recall, F1-Score
  + Benchmarking

### Deployment

### Abschlusspräsentation erstellen

### Dokumentation erstellen