# WBS

## Projektinitiierung

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

## Anforderungsanalyse

* funktionale Anforderungen spezifizieren
  + GUI
  + Einzulesende Datenformate
  + Akronym Expansion
  + Ausschreiben von Abkürzungen
* nichtfunktionale Anforderungen spezifizieren
  + Schnittstellen zu anderen Systeme: Was muss selbst programmiert werden, was kann selber implementiert werden?
  + Betriebsumgebung / Betriebssystem
  + Datenbanken
  + Datencharakteristika
* Technische Architektur
* Prozessbeschreibung
* Anforderungen priorisieren
* Glossar erstellen
* Pflichtenheft erstellen

## Datensammlung

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

## Explorative Datenanalyse / Datenqualität bestimmen

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

## Datenbereinigung

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

## Datenvorbereitung

* OCR (OpenVC, Pytesseract)
* Satzsegmentierung
* Worttokenizierung
* Normalisierung
  + stop word removal
  + stemming and lemmatization
  + removing digits/punctuation
  + lowercasing
* POS-Tagging
* 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

## Oberflächenmodellierung

## Deployment

## Abschlusspräsentation erstellen

## Dokumentation erstellen