Projektarbeit

für

Hui Chen

Matr.-Nr.: 10048521

Thema: Automatisierte Erfassung und Extraktion von Daten aus Tabellen in Dokumenten.

Aufgabenstellung:

Daten sind die Grundlage für Informationen und somit die Generierung von Wissen. Große Teile unserer Daten werden nach wie vor in Form von Tabellen innerhalb eines Dokuments publiziert, um eine spezifische These zu bestätigen oder zu widerlegen - ein Beispiel hierfür sind wissenschaftliche Veröffentlichungen. Neben der Verwendung für die ursprüngliche Fragestellung lassen sich Daten auch für andere Thesen nutzen. Dazu müssen sie jedoch zunächst extrahiert, normalisiert und klassifiziert werden.

Im Rahmen dieser Arbeit soll daher ermittelt werden, welche Technologien zur Extraktion von Daten bzw. Tabellen aus Dokumenten existieren. Im nächsten Schritt soll zu diesem Zweck ein Python-Skript entwickelt und getestet werden. Hierbei sind bestehende Algorithmen bzw. Programmbibliotheken zur Bildverarbeitung zu nutzen. Nach dem erfolgreichen Einsatz mit einem Beispieldokument soll das Skript auf die Stapelverarbeitung mehrerer Dokumente erweitert werden. Die extrahierten Daten sollen in eine Datenbank geladen werden.

Es ist abschließend zu diskutieren wie eine Normalisierung und Klassifizierung dieser Daten durchzuführen ist. Ziel ist es, Daten aus verschiedenen Dokumenten vergleichen zu können.

Die Ergebnisse der Arbeit sind in vorgegebener Form zu dokumentieren, anhand von Literaturdaten zu diskutieren und im Rahmen einer Abschlusspräsentation am Institut für Mehrphasenprozesse vorzustellen.

Beginn der Arbeit: 01.08.2022

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Institutsleitung Erstbetreuung: xyz Zweitbetreuung: Dr. H. Zernetsch

Der Zeitraum zur Bearbeitung der Projektarbeit beträgt 3 Monate ab dem Beginn der Arbeit. Für die Einhaltung der Bearbeitungszeit sowie des angefügten Zeitplans ist der Student/die Studentin selbst verantwortlich. Bei Abweichungen ist rechtzeitig Kontakt zur Betreuung zu suchen.

Zeitplan

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Monat 1 | | | | Monat 2 | | | | Monat 3 | | | |
| Recherche von bestehenden Technologien |  |  |  |  |  |  |  |  |  |  |  |  |
| Entwicklung des Skripts |  |  |  |  |  |  |  |  |  |  |  |  |
| Testung des Skripts und Optimierung |  |  |  |  |  |  |  |  |  |  |  |  |
| Ansätze zur Normalisierung und Klassifizierung |  |  |  |  |  |  |  |  |  |  |  |  |
| Schriftliche Ausarbeitung |  |  |  |  |  |  |  |  |  |  |  |  |