|  |  |
| --- | --- |
| Projektstatus Report | Status gesamt:  **grün** |

# Projekt: BibTexKonverter, Gruppe: David Konieczny, Lars Lafleur, Constantin Schulz

10.07.2024

|  |  |
| --- | --- |
| **Erfolge**, behobene Risiken oder Probleme (Farbcodierung = wie wichtig war Erledigung für Gesamterfolg) | * Erfolg 1 (wichtiger Milestone): Trotz Ausscheidens von Jürgen Bullinger konnte zum Durchstich ein funktionierender Prototyp erstellt werden * Erfolg 2 (wichtiger Milestone): Für die BibTex-Referenzen book, article, proceedings, inproceedings, incollection und phdthesis wurden jeweils 1000 Datensätze aus diversen Datenbanken zusammengetragen bzw. künstlich erzeugt, Aufteilung ca. 700 : 300 in Trainings- und Testdaten * Erfolg 3 (wichtiger Milestone): Als Weboberfläche dient Flask-Framework. Kopplung des Konvertierungsmodells mit der Weboberfläche erfolgt. String-Referenzen können bereits erkannt werden. * Erfolg 4 ( mittelmäßig wichtiger Milestone ): OCR mit Tesseract bereits implementiert; die Auswahl des Zeichensatzes für die jeweilige Sprache muss noch über die Benutzeroberfläche eingebunden werden * Erfolg 5 ( mittelmäßig wichtiger Milestone ): Erster Entwurf einer Evaluationsmetrik auf Basis des Vergleichs von Literaturtypen, Attributen und längsten gemeinsamen Teilfolgen * Erfolg 6 (wichtiger Milestone): Folgendes Felder können bereits erfolgreich extrahiert werden: authors, editors, year, volume, number, edition, pages, url, doi, month, address, publisher, school * Erfolg 7 (wichtiger Milestone): Wenn die DOI vorhanden ist, dann können mit ihr und der Python Bibliothek requests die Felder Publisher, Titel und Bocktitle ermittelt werden. Dies stellt zudem einen ersten wichtigen Schritt zur Vervollständigung der Einträge mithilfe Pybliometrics, Scholarly, Arxi und Pybliographer dar. |
| **Risiken und Probleme** (Farbcodierung = wie groß ist das Risiko für Gesamterfolg) | * Risiko 1 (mittleres Risiko ): Training des Modells auf Erkennung des BibTex-Referenzformats (article, book, etc.) noch nicht gestartet. Aufwand noch nicht abschätzbar, da zunächst die Erkennung der einzelnen Attribute im Vordergrund steht * Risiko 2 (mittleres Risiko ): Jürgen Bullinger ist definitiv nicht mehr im Projektteam |
| **Geplante nächste Aktivitäten** (Farbcodierung = wie wichtig ist Erledigung für Gesamterfolg) | * Aktivität 1 (wichtige Aktivität): Überprüfung der Qualität der Trainingsdaten (Sind die mit Pybtex erzeugten Literaturstring nach den jeweiligen Richtlinien des Styles formatiert?) * Aktivität 2 (wichtige Aktivität): Es soll für die Typen book, article, proceedings, inproceedings und incollection jeweils ein binärer Klassifizierer trainiert werde. Jeder Klassifizierer spezialisiert sich auf das Erkennen von einem Typen. * Aktivität 3 (wichtiger Aktivität): Verwendung der generierten Trainingsdaten für das ML-Modell: Liefert das Modell anhand der selbst erzeugten Trainingsdaten zufriedenstellende Ergebnisse? * Aktivität 4 (wichtige Aktivität): Extrahierung der Felder title, booktitle, series und journal * Aktivität 5 (mittelmäßig wichtige Aktivität): Testen der Bibliotheken zum Dursuchen von Litertaturdatenbanken zur Auflösung von Abkürzungen: Pybliometrics, Scholarly, Arxiv, Pybliographer * Aktivität 6 (wichtiger Aktivität): OCR für mehrere Sprachen einrichten, vorteilhaft wäre ein latin-Zeichensatz, alternativ: Implementierung eines Dropdowns in Weboberfläche für Vorgabe der Sprache * Aktivität 7 ( mittelmäßig wichtige Aktivität ): Anpassung der Weboberfläche durch erklärende Tooltips (Anregung aus Peer-Feedback) * Aktivität 8 ( mittelmäßig wichtige Aktivität ): Optimierung der virtuellen Entwicklungsumgebung zur Installation des Projektes * Aktivität 9 ( mittelmäßig wichtige Aktivität ): Feintuning der Evaluationsmetrik, z. B. detaillierter Attributvergleich von Pflichtattributen pro Literaturtyp, da diese mehr ins Gewicht fallen * Aktivität 10 ( mittelmäßig wichtige Aktivität ): Evaluierung der Ergebnisse des Modells bei Anwendung auf Testdaten und Visualisierung der Evaluationsergebnisse * Aktivität 11 ( mittelmäßig wichtige Aktivität ): Beginn mit Dokumentation der bisherigen Methoden und Prozesse |
| Offene Fragen, generelle Anmerkungen | In welcher Form soll das finale Produkt vorliegen? Reicht ein zu klonendes Repository mit auszuführender main.py aus?  Feedback der Gruppe Abschlussarbeiten verweist auf Aussagen zu wünschenswerter Markdown Syntax. Gruppe wurde bereits angeschrieben, bis jetzt aber keine Reaktion.  Anregung bzgl. Einlesen ganzer Bibliographien: In welchem Format liegen diese vor? Sind damit Screenshots von ganzen Literaturverzeichnissen, z. B. vom Ende eines Papers gemeint? |